Computer-aided Cytological Grading Systems for Fine Needle Aspiration Biopsies of Breast Cancer

Muneera Alsaedi

A Thesis
in
The Department
of
Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements
For the Degree of
Doctor of Philosophy (Computer Science) at
Concordia University
Montréal, Québec, Canada

May 2019

Concordia University

School of Graduate Studies

This is to certify that the thesis prepared

By: **Mrs. Muneera Alsaedi**

Entitled: **Computer-aided Cytological Grading Systems for Fine Needle Aspiration Biopsies of Breast Cancer**

and submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Computer Science)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____ Chair
Dr. Abdel R. Sebak

_____ External Examiner
Dr. Catherine Laporte

_____ Examiner
Dr. Maria A. Amer

_____ Examiner
Dr. Tristan Glatard

_____ Examiner
Dr. Ching Suen

_____ Supervisor
Dr. Thomas Fevens

_____ Co-supervisor
Dr. Adam Krzyżak

Approved by _____
Dr Volker Haarslev, Graduate Program Director

31 July 2019 _____
Dr Amir Asif, Dean
Gina Cody School of Engineering and Computer Science

Abstract

Computer-aided Cytological Grading Systems for Fine Needle Aspiration Biopsies of Breast Cancer

Muneera Alsaedi, Ph.D.

Concordia University, 2019

According to the American Cancer Society, breast cancer is the world's most commonly diagnosed and deadliest form of cancer in women. A major determinant of the survival rate in breast cancer patients are the accuracy and speed of the malignancy grade determination. This thesis considers the classification problem related to determining the grade of a malignant tumor accurately and efficiently. A Fine Needle Aspiration (FNA) biopsy is a key mechanism for breast cancer diagnosis as well as for assigning grades to malignant cases. Carrying out a manual examination of FNA demands substantial work from the pathologist which may result in delays, human errors, and consequently lead to misclassified grades. In this context, the most common grading system for microscopic imaging for breast cancer is the Bloom and Richardson (BR) histological grading system which is based on the evaluation of tissues and cells. BR is not directly applicable to FNA biopsy slides due to distortion of tissue and even cell structures on the cytological slides. Therefore, in this thesis, to grade FNA images of breast cancer, instead of the BR grading scheme, six known cytological grading schemes, three newly proposed cytological grading schemes, and five grading systems based on convolutional neural networks were proposed to automatically determine the malignancy grade of breast cancer.

First, considering traditional Machine Learning methods, six cytological grading systems (CA-CGSs) based on six cytological schemes used by pathologists for FNA biopsies of breast cancer were proposed to grade tumors. Each system was built using the cytological criteria as proposed in the original CGSs. The six considered cytological grading schemes in this thesis were Fisher's modification of Black's nuclear grading, Mouriquand's grading, Robinson's grading, Taniguchi et al's, Khan et al's and Howell's modification in mitosis count criteria. To fulfill this task, different sets of handcrafted features using customized image processing algorithms were extracted for classification purpose. The proposed systems were able efficiently to classify FNA slides into G2 (moderately malignant) or G3 (highly malignant) cases using traditional machine learning algorithms. Additionally, three new cytological grading systems were proposed by augmenting three of the original CGSs by adding the low magnification features. However, the systems were not sensitive enough with regards to G3

cases due to the low number of available data samples. Therefore, a data balancing was performed to improve the sensitivity for G3 cases. Consequently, in the second objective of this work, data sampling and RUSBoost methods were applied to the datasets to adjust the class distribution and boost the sensitivity performance of the proposed systems. This enabled a sensitivity improvement of up to 30% which highlights the significance of class balancing in the task of malignancy grading of breast cancer.

Additionally, due to the considerable time and efforts required for handcrafted features-based cytological grading systems in order to achieve efficient feature engineering results, a deep learning (DL) approach was proposed to avoid the aforementioned challenges without compromising the grading accuracy. Thus, in this thesis, five different pre-trained convolutional neural network (CNN) models, namely GoogleNet Inception-v3, AlexNet, ResNet18, ResNet50, and ResNet101, combined with different techniques to deal with unbalanced data, were used to develop automated computer-aided cytological malignancy grading systems (CNN-CMGSs). According to the obtained results, the proposed CNN-CMGS based on GoogleNet Inception-v3 combined with the oversampling method provides the best accuracy performance for the problem at hand.

The results demonstrated that the proposed CGSs are highly correlated since they share some of the cytological criteria. Further, the overall accuracy of the CGSs is roughly the same and overall, the handcrafted features-based CGSs performed best even in the absence of class distribution rebalancing. Overall, for case classification, the best results were obtained for computer-aided CGSs based on the modified Khan *et al.*'s and Robinson's schemes with accuracies of 97.77% and 97.28%, respectively. Meanwhile, for patient classification, the overall best results were obtained for computer-aided CGSs based on the modified Khan *et al.*'s and modified Fisher's schemes with accuracies of 96.50% and 95.71%, respectively. These results surpass previously reported results in the literature for computer-aided CGS based on BR histologic grading. Moreover, in clinical practice, Robinson's typically has the best diagnostic accuracy with the highest reported experimental accuracy rate of 90%. Thus, the obtained results demonstrate that computer-aided breast cancer cytological grading systems using FNA can potentially achieve accuracy rates comparable to the more invasive histopathological BR-method.

Acknowledgments

First and foremost, praises and thanks to God for assisting me to complete the research successfully. I am so grateful to you, Allah, for seeing me through the trials and tribulations of the long, arduous, yet rewarding journey.

I would like to acknowledge my government, specifically, the Iraqi Commission for Computers and Informatics, Informatics Institute for Postgraduate Studies in Iraq for financially supporting my research, and granting me the opportunity to study abroad.

I would like to acknowledge everyone who played a role in my academic accomplishments. First, I would like to express my deep and sincere gratitude to my research supervisors, Dr. Thomas Fevens and Dr. Adam Krzyzak, for giving me the opportunity to do my work and providing invaluable guidance throughout this research. Without your guidance and feedback, this work would not have been possible.

Also, deep gratitude to my committee members, each of whom has provided patient advice and guidance throughout the research process.

I would like to thank the Department of Computer Science and Software Engineering at Concordia University, specifically, Program Advisor Halina Monkiewicz for her full assistance.

I would like to thank my entire family, and particularly, my dear mother and my younger sister Zahra. Thank you for your love, support, prayers, and understanding. Without your support, I could never have reached this current level of success.

Thank you to a very special friend who has supported me through this journey.

Finally, my sincere thanks to each and every person who has supported me through this research, directly or indirectly.

Dedication

To the spirit of my dear father, you will always be with me.

Contents

List of Figures

List of Tables

List of Abbreviations

AdaSS	**A**daptive **S**plitting and **S**election
ANN	**A**rtifical **N**eural **N**etwork
BR	**B**loom **R**ichardson **G**rading **S**cheme
BN	**B**atch **N**ormalization
CGSs	**C**ytological **G**rading **S**ystems
CN	**C**ell **N**ecrosis
CFS	**C**orrelation-based **F**eature **S**election
CNN	**C**onvolutional **N**eural **N**etworks
CMG	**C**ytological **M**alignancy **G**rading
CHT	**C**ircle **H**ough **T**ransform
DL	**D**eep **L**earning
DFS	**D**isease **F**ree **S**urvival
DR	**D**imensionality **R**eduction
DT	**D**ecision **T**ree
DT-AdaBoost	**A**daptive **B**oosting of **D**ecision **T**rees
FNA	**F**ine **N**eedle **A**spiration
FCBF	**F**ast **C**orrelation-based **F**eature selection
FFNN	**F**eedfoward **N**eural **N**etworks
FP	**F**alse **P**ositive
FN	**F**alse **N**egative
FCM	**F**uzzy **C**-**M**eans
GLCM	**G**ray **L**evel **C**o-occurrence **M**atrix
GLRLM	**G**ray **L**evel **R**un **L**ength **M**atrix
GVF-snake	**G**radient **V**ector **F**low-snake
GVF-MO	**G**radient **v**ector **F**low-snake-**M**orphological **O**peration
HMI	**H**igh **M**agnification **I**mages

HMG	**H**igh **M**alignancy **G**rade
HPF	**H**igh **P**ower **F**eld
HT	**H**ough **T**ransform
HSV	**H**ue **S**aturation **V**alue
IMG	**I**ntermediate **M**alignancy **G**rade
ILSVRC	**I**mageNet **L**arge **S**cale **V**isual **R**ecognition **C**ompetition
KNN	**K**-**N**earest **N**eighbors
LMI	**L**ow **M**agnification **I**mages
LMG	**L**ow **M**alignancy **G**rade
LDA	**L**linear **D**discriminant **A**analysis
LLE	**L**ocally **L**inear **E**mbedding
LEM	**L**aplacian **E**igen **M**aps
LS	**L**evel **S**et
Lab	**L**ightness value **G**reen-**R**ed value **B**lue-**Y**ellow value
LUV	**L**ightness **UV**-chromaticity
MC	**M**itosis **C**ount
MBF	**M**arkov **B**lanket **F**ilter
MLP	**M**ultilayer **P**erceptron
MO	**M**orphological **O**perations
NTN	**N**aked **T**umor **C**ell **N**uclei
NCR	**N**uclear **C**ytoplasm **R**atio
NBR	**N**ottingham **B**loom **R**ichardson
NG	**N**ulcear **G**rading
NB	**N**aive **B**ayes
NS	**N**uclear **S**ize
OSD	**O**versample **D**ata
OS	**O**verall **S**urvival
OD	**O**ptical **D**ensity
PCA	**P**rincipal **C**omponents **A**nalysis
PDF	**P**robability **D**ensity **F**unction
PSO	**P**article **S**warm **O**ptimization
RUSBoost	**R**random **U**ndersample **B**oosting
RGB	**R**ed **G**reen **B**lue
RST	**R**otation **S**caling **T**ranslation
RFDT	**R**andom **F**orest of **D**ecision**T**rees

ReLU	**R**ectified **L**inear **U**nit
ResNet	**R**esidual **N**etwork
ROC	**R**eceiver **O**perating **C**haracteristic
RGB	**R**ed **G**reen **B**lue color space of the image
SFS	**S**equential **F**orward **S**election
SBE	**S**equential **B**ackward **E**limination
SVM	**S**upport **V**ector **M**achines
SBR	**S**carff **B**loom **R**ichardson
SOM	**S**elf **O**rganizing **M**aps
TP	**T**rue **P**ositive
TN	**T**rue **N**egative
USD	**U**ndersample **D**ata

Chapter 1

Introduction

As one of the deadliest forms of cancer with an alarming rate of mortality, breast cancer certainly represents a threat to all adult women. According to the American Cancer Society [1], in 2019, an estimated 268,600 new cases of invasive breast cancer were expected to be diagnosed among women in the United States (US) along with about 2,670 new cases expected in men. Breast cancer is the leading cause of cancer death among women aged 20 to 59 years in the US. Further, in 2018, breast cancer was the world's most commonly diagnosed and deadliest form of cancer among women. Also, approximately 40,610 women and 460 men were expected to die from breast cancer in 2017 [1]. Moreover, from 2005 to 2014, the rate of invasive and in situ breast cancer cases lightly increased among women over the age of 50 (see Figure 1). This rate has been steadily increasing by 0.2% yearly since the mid-1990s among women under the age of 50, as shown in Figure 1. On the other hand, there are discrepancies between rich and poor countries in terms of treatment facilities and even a lack of early detection. Concretely, this means that approximately 90 per 100,000 women in Eastern Europe and 30 per 100,000 women in Eastern Africa are diagnosed with breast cancer annually. Among them, about 15 per 100,000 do not survive the disease [2]. In medical terms, breast cancer is defined as "a malignant tumor that starts in the cells of the breast. A malignant tumor is a group of cancer cells that can grow into (invade) surrounding tissues or spread (metastasize) to distant areas of the body" [3]. To increase the chances of recovery, breast cancer has to be detected and treated in its very early stages. Indeed, the effectiveness of the treatment largely depends on the timely detection and precise grading of the disease.

Over the years, doctors and pathologists have efficiently used Fine Needle Aspiration (FNA) slides for the diagnosis of breast lesions [4]. Carrying out a manual examination of FNA is associated with

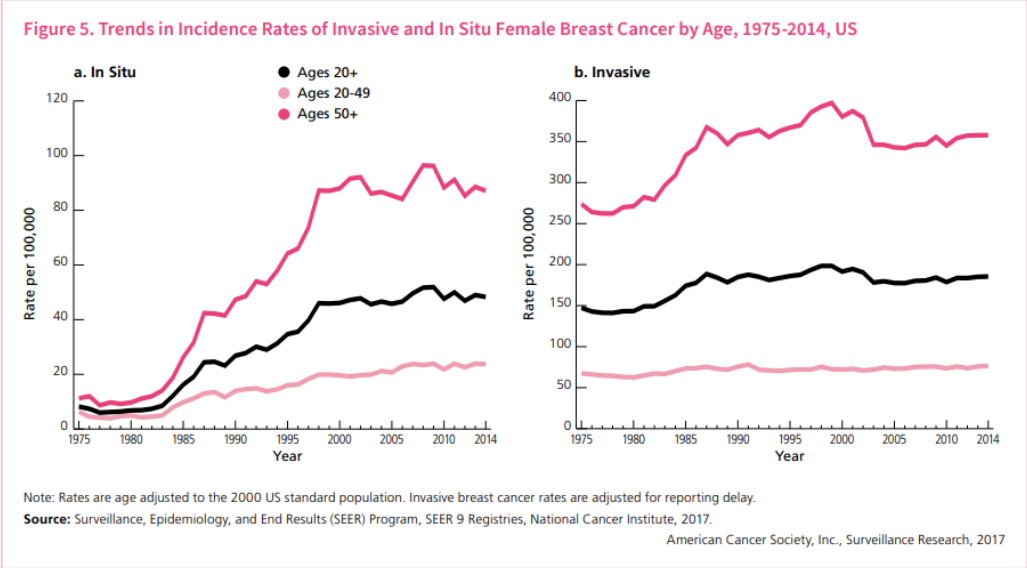


Figure 1: Rates for invasive and in situ breast cancer cases per age at diagnosis [1].

a substantial amount of work to the pathologists (an average of 80 slides per day, having thousands of cells each) [5], which is a challenging task and can lead to missed or delayed, and consequently, misclassified grades due to excessive fatigue. Further, according to current pathology-based studies, two significant types of errors have a major impact on breast cancer grading. The first type is related to pathologic misinterpretation errors that ranged from 5 to 50.7% [6]. The second type is due to the nature of breast cancer lesion and can go up to 68% [7]. Therefore, due to the mentioned limitations, it is of paramount importance to provide an expert second opinion or mechanism to verify the accuracy of the grade provided by the FNA manual exam. This is particularly important for the cases where pathologists required further manual examinations for FNA. For example, it helps to avoid over-treatment for low malignancy grade patients and allows screening on a large scale to identify systematic errors.

1.1 The problem

The classification problem related to determining the grade of a malignant tumor accurately and in a timely manner is the research problem we seek to address in this thesis. It is widely known that early stage tumors can be treated. While it is widely known that treating advanced stage tumors is difficult, early stage tumors can be more easily treated; thus, early detection can save lives. For early-stage treatment, mastectomy has now been replaced by breast-conserving surgery (lumpectomy) followed by local radiotherapy. Other treatments given to advanced stage patients include combinations of chemotherapy and hormonal therapy [8]. The aim of this work is to propose

an automatic cytological malignancy grading system for FNA biopsies of breast cancer. Pathologists, on a daily basis, go through hundreds of cytological images for grading breast cancer. Grading a large number of images with a huge amount of cells is both challenging and tiring. In addition, in some cases, further examination is required by a specialist. In this context, human errors can occur which can threaten the survival of a patient. Thus, there is a significant need for a computer-aided malignancy grading system for FNA biopsies which not only accurately determines the cancer grade but also assists pathologists by providing a second opinion in difficult or challenging situations. In this thesis, a classification of the malignancy is proposed using cytological grading schemes which, to the best of the author's knowledge, have never been investigated in the area of computer science. The following section discusses the intended approach to solve this problem.

1.2 The approach

The problem of classifying malignancy of breast cancer is approached via three objectives which are listed as follows:

1. **First objective:** Six computer-aided cytological grading systems (CA-CGSs) for FNA biopsy of breast cancer, each having its own scheme with criteria that describe the malignancy level of the cancer cells, were first proposed using traditional classification systems. These systems take cytological images as input, and output the grade (G2 or G3) of the tumor. The images used in this thesis are from two magnification levels (1) High (400X) and (2) Low (100X). Basically, three of these systems use only the high magnification images and the other three systems use both low and high to evaluate and assign a malignancy grade for FNA biopsies. Thus, as the second stage of this objective, new original modified versions of three of the six CGSs that were used only the high magnification images were proposed. Several fundamental stages were applied to the images to segment them and then to estimate specific cellular and nuclear properties for the classification purpose. The traditional approach of classification, which requires a lot of manual tasks and is time-consuming, was used. Moreover, since the distribution of the classes in the used dataset was not uniform, automatic classification of the cancer grade was compromised for low data samples as compared to high data samples. This kind of problem is generally called an imbalanced classification problem.

2. **Second objective:**The aim is to solve the imbalanced classes distribution problem of the used datasets in this research. As mentioned earlier, imbalanced dataset create classification errors for the data with fewer samples often referred to as minority class. This induces a reduction in the performance accuracy of the classification models on the minority class predictions.

Three data balancing techniques have been used in this thesis to solve this problem, namely (1) Oversampling, (2) Undersampling, and (3) Hybrid RUSBoost. The oversampling method duplicates the data of the minority class such that the majority and minority classes both have the same number of samples and consequently, the definition of majority and minority class is no longer valid using this technique. In contrast, in the undersampling technique, only a few randomly selected data samples from the majority class are used in order to equate both classes. The hybrid RUSBoost is a random undersampling combined with AdaBoost classifier where, in contrast to other data sampling techniques, the boosting is used for weak learners. The application of these techniques not only improved the classification performance for the minority class predictions but also the overall accuracy of the system.

3. **Third objective:** Despite the high performance achieved by CNN models for the malignancy diagnosis of breast cancer using different imaging modalities, the malignancy grading problem for cytological images of breast cancer has been much less studied using CNN models. Further, traditional classification systems heavily rely on accurate image segmentation and handcrafted feature extraction tasks which are time-consuming and require domain expertise. Thus, considering the aforementioned challenges, we focused on proposing the first CNN-based computer-aided cytological malignancy grading system (CNN-CMGS) for cytological images of FNA biopsies of breast cancer to address the gap in the available literature and to facilitate the mentioned difficulties and challenging tasks associated with traditional classification systems. The convolutional neural network models are a combination of filters and functions that help automatically classify the FNA biopsy images of breast cancer without all the aforementioned pre-processing and segmentation stages. The results obtained from the CNN models are promising and less time-consuming as compared to traditional classification systems. The only trade-off which was found in this thesis was the overall accuracy of the CNN-based CMGS for the patient classification task, which is 4% more than handcrafted features-based CGSs.

1.3 Main contributions

In this thesis, six CGSs for assigning the malignancy grades of breast cancer were proposed. The classification involves traditional approaches including image pre-processing, data balancing techniques, and lastly, CNN models to predict the malignancy grades of cancer in less time and with an accuracy comparable to that of the traditional approaches. The main contributions of this work are as follows:

1. Proposed nine novel computer-aided CGSs for FNA biopsies of breast cancer based on pathology-guided handcrafted features.

2. Introduced a novel method for the estimation of three nuclear features; namely, the mitosis count (MCC), naked tumor nuclei (NTCN), and nuclear-cytoplasmic ratio (NCR) cytological characteristics of cytological systems.

3. Proposed the first CNN-based CA-CMGS for breast cancer along with the adaptation of data sampling with CNN models.

4. Analyzed different data sampling techniques and demonstrated the superiority of the oversampling data (OS) for both traditional and CNN-based classification models.

All of the above contributions are described in the subsequent chapters of this thesis.

1.4 Research publications

The work in this research has been published in the following papers:

- Alsaedi, M., Fevens, T., Krzyżak, A., & Jeleń, Ł. (2017, November). Cytological malignancy grading systems for fine needle aspiration biopsies of breast cancer. In 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 705-709). IEEE.

- Alsaedi, M., Fevens, T., Krzyżak, A., & Jeleń, Ł. (2018, May). Hybrid RUSBoost Versus Data Sampling to Address Data Imbalance for Breast Cancer Cytological Malignancy Grading. In 2018 International Conference on Pattern Recognition and Artificial Intelligence (ICPRAI) (pp. 545–551).

1.5 Structure of the thesis

This thesis consists of six chapters as follows:

Chapter 2 presents breast cancer from a medical point of view and reviews the literature related to computer-aided diagnosis and grading of breast cancer.

Chapter 3 introduces six handcraft-based CGS systems for cytological malignancy grading. This chapter also includes several steps for image pre-processing and traditional machine learning classification. Furthermore, three modified CGSs based on low magnification features are introduced.

Chapter 4 applies imbalanced data techniques to adjust the class distribution and to boost the performance of classification systems.

Chapter 5 introduces the first CNN-based cytological malignancy grading system for FNA biopsies of breast cancer with different data types.

Chapter 6 presents conclusions based on the findings and addresses significant tasks for future research.

Chapter 2

Background and literature review

Breast cancer is posing a real threat to women across the world. According to the International Agency for Research on Cancer (2018) [9], nowadays, breast cancer is the second most common and deadliest cancer. Also, among females, it is the most commonly diagnosed cancer and the leading cause of cancer death, followed by colorectal and lung cancer. Further, according to the last statistics of 2017, [9] it was anticipated that there will be about 2.1 million newly diagnosed female breast cancer cases in 2018, accounting for almost 1 in 4 cancer cases among women.

Breast cancer can be described as a heterogeneous tumor resulting in a complex disease with different morphological and biological characteristics. The intra and inter tumor diversity of breast cancer is high. Patients with breast cancer have internal diversities as well. Breast cancer management heavily relies on the availability of efficient clinical and pathological prognostic and predictive factors. These factors influence diagnosis. According to medical researchers, to significantly diminish breast cancer mortality rates, early, proper, and accurate diagnosis of the disease is required. Typically, two major obstacles in overcoming breast cancer are its detection time and the accuracy with which the malignancy degree is determined. A substantial amount of research works have focused on breast cancer detection, either medically or using computer simulations. However, it is impossible to start breast cancer therapy unless the malignancy grade is precisely determined. Therefore, an important step is to accurately grade tumors so that the most appropriate medical regimen is selected, which requires extensive knowledge and experience from the pathologist responsible for the diagnosis.

2.1 Pathology-based diagnosis of breast cancer

A prime reason for the increase in the number of global deaths due to breast cancer is the lack of proper diagnosis at very early stages. Medical examinations are of utmost importance in this case. An important and often used diagnosis method for breast cancer is the so-called triple-medical test which is based on three medical examinations and is used to achieve high confidence in the diagnosis. It includes breast self-examination (called "palpation", carried out by the patients themselves), screening examinations (mammography or ultrasound) and Fine Needle Aspiration (FNA) biopsy (abnormal tissue collection from a suspicious area by a syringe with a fine needle) [10]. A biopsy is a medical procedure that removes a part or the integrity of a tumor. In this context, FNA has its own pros and cons. To begin with, FNA does not require the skin to be cut (excised biopsy from a tumor). Moreover, the FNA procedure is not time-consuming. Normally it is possible to diagnose the patient on the same day. Compared to other histological approaches (tissue excision) carried out during surgery, FNA is far less invasive and inexpensive. In cases which it is properly performed and a high quality cytopathology service is available, FNA is proven to be safe, simple, fast, and cost-effective [11]. However, this technique also suffers from several limitations. For instance, when using FNA, there is a high probability of sampling error occurrence due to the wrong position of the sample. Additionally, FNA heavily relies on the breast cytopathology expertise of the pathologist, making it a pathologist-dependent method. There is also frequently a lack of reliable histological architecture (tissue structure). Finally, since FNA involves the extraction of material from the patient's body, it is very important that the extraction is done accurately. Therefore, the extraction itself is a challenge that further complicates the FNA. Indeed, a marginal error in the material extraction can result in the destruction of the tissue structure, leading to difficulties in evaluating the tumor features precisely. Thus, the procedure has to be undergone in a thoroughly professional manner, to ensure that the patient is not put under inappropriate treatment [11].

Furthermore, the FNA test also has two potential diagnosis downsides as a consequence of which an additional imaging or cytopathological assessment is required. One of the downsides is that, as mentioned above, extraction should be done accurately; if not, the extracted tissue material can be diagnosed as normal cells when they are indeed cancerous cells. This leads to inaccurate diagnosis, known as a false negative result of the FNA examination. The false negative result might cause a missed or delayed diagnosis of breast cancer which decreases the survival chances of the patient [12]. The FNA test presents a 2% to 4% false negative rate which is significantly lower than other tests such as the mammogram, where about 6% to 46% of women with invasive cancer have a false negative mammogram, usually more common in younger women, with dense breasts, and who have mucinous, lobular, or rapidly growing cancers [13, 12]. Further, FNA may provide limited

information about the breast cancer type identification. More precisely, they cannot determine if the cancer is invasive or non-invasive (ductal carcinoma or in situ) [12]. The other downside of the FNA biopsy test is that the extracted tissue material can be diagnosed as cancerous cells when they are indeed normal cells. This causes a false positive result and can range between 0.2 and 0.3% [14].

Among the aforementioned three tests, FNA biopsy is the most important and is performed as a pre-operative test to evaluate breast lumps [15]. It can also sample different areas of the lesion, in contrast to core needle biopsy, which is a method used to take a small amount of suspicious tissue from the breast using a larger needle [16, 17]. There are other types of biopsies, such as core-needle biopsy which, depending on the case, involves the removal of either a small lesion portion or the entire , and surgical or excision biopsy, which is based on the removal of either the entire tumor or a small part of it during surgery. The doctor's judgment of the appropriate biopsy type to be performed depends on various factors. These factors involve the appearance, size, and location of the suspicious area in the breast. Figure 2 shows a graphical representation of the different biopsy types.

The key feature of the FNA biopsy is that it is used by pathologists to distinguish malignant and benign tumors and to assign grades to malignant cases. In particular, it allows the classification of tumors, therefore providing specific and suitable treatment to the patient as early as possible. Early diagnosis is life-saving, cost-effective, and requires less aggressive therapy. The classifications include malignancy diagnosis and malignancy grading problems.

2.1.1 Diagnosis of breast cancer

Once an FNA is carried out, the collected specimen is put on a glass slide and stained. The type of staining depends on the type of cell structures to be visualized. When the specimen is stained, a microscopic examination is performed to detect the presence of malignancy in a tumor. Once the breast cancer is confirmed, the pathologist provides more cancer details (type, malignancy grade, and malignancy stage) which will be later combined with specific factors (prognostic and predictive) to determine the progression of the case. Predictive factors enable the prediction of the undertaken treatment while prognostic factors allow for the prognosis of the overall survival (OS) and disease-free survival (DFS) rates [2, 10].

2.1.2 Malignancy grading of breast cancer

Once a tumor is diagnosed as malignant, the issue becomes more complex. Determining the malignancy degree (grading) is of paramount importance at this point since, depending on the grade of the tumor, proper prognostic treatment should be provided to the patient in a timely manner. The

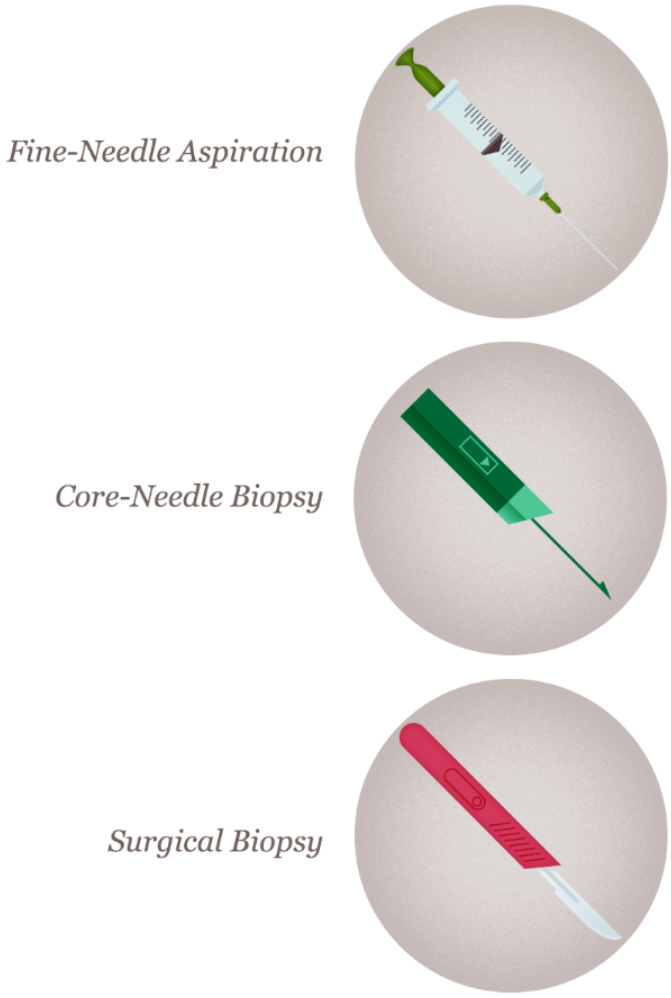


Figure 2: Biopsy types [17] .

key objective of this study is to investigate the cytological grading systems of breast cancer with the aim of proposing a computer-aided cytological grading framework for breast cancer using the cytological data of the FNA slide.

What is the grade of a breast tumor? How is it determined?

The tumor grade provides a description of a malignant tumor in terms of how its tissue and cells look under a microscope to a pathologist. Hence, it indicates the likelihood of tumor cells to grow and spread. Consequently, to classify tumor cells, grading systems or schemes are used. Depending on the type of cancer, these schemes vary. Generally, malignant tumors can be graded as 1, 2, or 3, depending on the degree of abnormalities in the cancerous cells and their nuclei. However, if the grading system is not specified for a certain cancer type, pathologists usually use a grading system

that consists of the following tumor classification [18]:

- GX: Grade cannot be determined.
- G1: Well differentiated cells (assigned as low grade).
- G2: Moderately differentiated cells (assigned as intermediate grade).
- G3: Poorly differentiated cells(assigned as high grade).
- G4: Undifferentiated cells (assigned as high advantage grade).

If the cancer type is specified, such as breast cancer, prostate cancer, blood cancer, etc., pathologists usually use a specific grading system according to the type of cancer. Therefore, in cytological grading systems of breast cancer (research problem addressed in this thesis), malignant tumors can be graded as G1, G2, or G3. In G1, the appearance and arrangement of tumor cells within the slide look similar to normal cells. These tumors grow and spread slowly. Meanwhile, the appearance and arrangement of cell nuclei in G2 and G3 tumors are different from normal cells. Thus, G2 and G3 tumors tend to grow very quickly and spread faster than lower grade tumors [18]. The three major categories of malignancy grades and the corresponding medical interpretation that are used to help select and manage the most appropriate remedy for each breast cancer case are [18]:

- G1: Cancerous cells are well differentiated w.r.t healthy cells (allocated to low grade).
- G2: Cancerous cells are moderately differentiated w.r.t healthy cells (allocated to intermediate grade).
- G3: Cancerous cells are poorly differentiated w.r.t healthy cells (allocated to high grade).

As mentioned earlier, the purpose of malignancy grading is to select the most suitable treatment for the patient. Generally, two types of grading systems are used by pathologists to grade the malignancy of breast tumors, namely the histological and the cytological grading of breast cancer.

Histological grading systems:

Several schemes are used to grade breast cancer histologically. However, the most widely used systems for breast tumor grading are based on the Bloom Richardson (BR) histological grading scheme [19]. In 1957, Bloom and Richardson came up with an innovative approach to grade breast tissue histologically. The survival rate from breast cancer is strongly correlated to the histological grade [20]. This system is particularly useful for histological images (see Figure 3) of thin slices of surgically excised tumor biopsies where the tissue structure is mostly preserved allowing for a

determination of tubule formations and cell nest structures (maintain or preserve the tissue and cell structures).

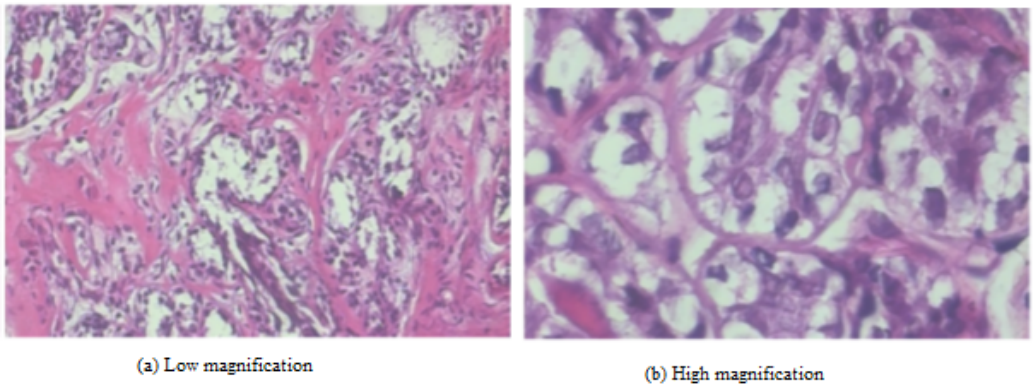


Figure 3: Example images of one case, (a) Low magnification (100x) and (b) High magnification (400x), from [21] .

In the BR system, three scales (1,2 or 3) are considered to evaluate the three malignancy factors: Tubules Formation, Cell Pleomorphism and MCC. Tubules Formation describes the presence of a tubular arrangement of the cells, Cell Pleomorphism denotes variations in the size, shape and staining of the nuclei, and MCC characterizes the mitotic activity of cells per their life. The malignancy of the tumor is assigned a grade which is based on the summation of the three scales of the aforementioned malignant factors and therefore, tumor malignancy is classified as grade G1, G2 or G3 [19]. The Nottingham Bloom Richardson (NBR) scheme is another widely used system to grade breast cancer histologically. The scheme is a modification of the BR scheme which includes a malignant indicator to check the involvement of lymph nodes in the tumor (metastasis) [22, 23, 24]. Histological grading, in spite of having the inherent capability of being able to predict overall and metastasis-free survival (local and regional breast cancer), suffers from potential limitations. Because of the irregularities in the cells' shapes, determining the level of malignancy consumes effort and time, due to inter and intra observation variations [25]. Analysis of histological images is not usually performed in-depth. Thus, significant malignancy information related to the internal cell structures (such as nuclei and their nucleolus) can be obtained by the deep cytology evaluation for cytological images of the FNA slides [26].

Cytological grading systems:

Due to the aforementioned difficulties associated with histological grading, cytological grading (cytology means the cells study in terms of structure, function and chemistry) using FNA biopsies was developed. The technique of FNA biopsy was first described in 1847 and was introduced in clinical practice by Ellis and Martin in the 1930s [27]. In recent years, FNA biopsy material is increasingly

being used in numerous centers in North America for pre-operative diagnosis of breast cancer [28]. Attempts have been made to identify various prognostic parameters or grading on FNA material to determine the best therapy for patients with breast cancer [29]. For an FNA biopsy, the material is extracted by a needle and smeared, usually non-uniformly, on a glass to create a slide for cytological imaging. This may result in the partial destruction of the tissue structure, and sometimes even of the nuclei of cells. Since this loss of tissue structure complicates the determination of tubule formation, combined with difficulties in scoring mitoses [30], cytological grading schemes, based on cell nuclei characteristics such as architecture, nuclear chromatin, dissociation of cells, nuclear features, etc. (see Figure 4), have been proposed instead of histologic schemes like the BR. These features display the aggressiveness of cancerous cells and utilize them to allocate grades for malignant tumors.

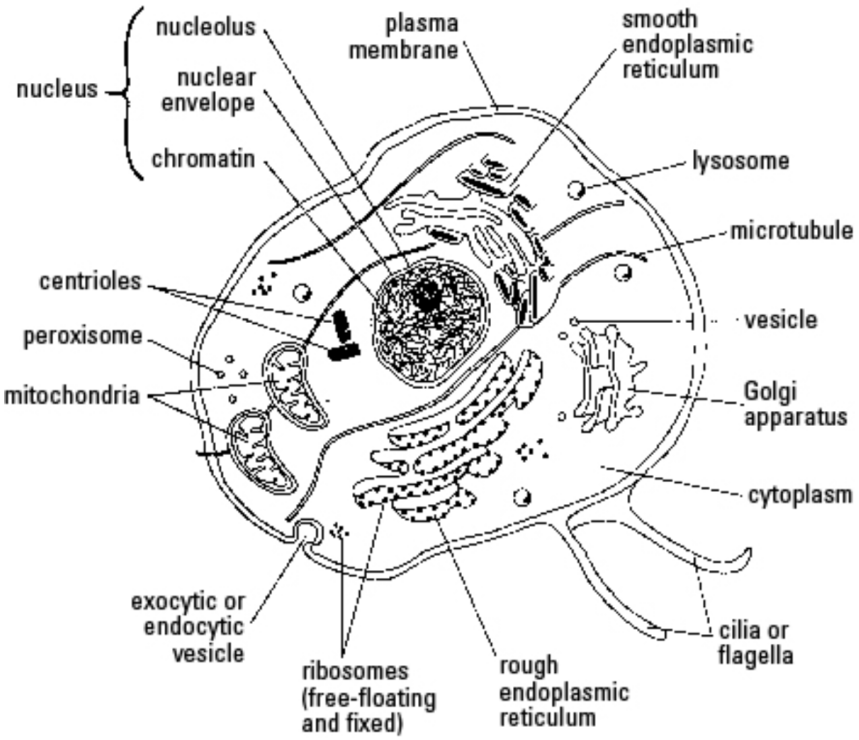


Figure 4: The general organization of a typical cell taken from [31].

Different studies have been conducted to evaluate cytological grading schemes for FNA biopsy slides with the aim of determining which system corresponds best to the BR-histological grading scheme and can be accepted by pathologists as a gold standard [32, 33, 29]. However, finding a solution to this problem is quite challenging due to the fact that the FNA slides are of low-resolution and are distorted by different medical stains. Therefore, the dependence on pathologists is drastically increased. In this thesis, to overcome the aforementioned difficulties, six computer-aided cytologic grading systems for FNA biopsies are proposed, each system is tailored to follow the cytological

characteristics defined by each of the published six cytological grading schemes used by pathologists.

The nuclear characteristics derived from FNA slides, specifically for the malignancy grading problem, refer to the level of cell mutation; therefore, cytological grading uses NG to a large extent. Examples of considered nucleus characteristics are the size and shape of the nucleus, the number and size of nucleoli, and the chromatin clumping which are used by Fisher's modification of Black's nuclear grading [34], Mouriquand's grading [35], Robinson's grading [36], and others [37, 38, 39]. Table 1 lists the considered cytological grading schemes together with their associated cytological characteristics. For instance, to grade cancer,the cytological characteristics for a typical scheme shown in Table 1 is evaluated according to the corresponding scoring mechanism defined in Table 2. This table summarizes the scoring methods of the aforementioned schemes. The scoring mechanism leads to a final malignancy grade which is determined by adding the scores of the malignancy characteristics associated with the considered cytological scheme.

To give more details about the six considered cytological grading schemes and their associated cytological criteria we discuss below the schemes as reported in the pathology-based studies. According to [40, 41], four systems (Robinson's [36], Howell's [37], khan *et al.* [38], and Taniguchi *et al.* [39]) out of the six mentioned schemes have used an objective scoring with categorical order for each criteria. Fisher's system [34] has no scoring method to score the criteria whereas Mouriquand's [35] system has a scoring method but not in the same categorical order.

To start, Fisher's system [34] does not use scoring methods like other systems and is a modification of Black's nuclear grading (NG) which studies the evaluation of the nuclear characteristics. Anisonucleosis, nuclear membrane, nuclear chromasia, nucleoli, chromatin pattern and mitoses count are the six criteria used by Fisher's methods to grade the cancers cytologically as NG G1, G2, and G3.

Mouriquand's system [35] gave a score of 0–3 for four criteria namely cellular characteristics, nuclear features, nucleolus, and mitotic figures. Based on this method, cancers are graded G1 if the total score is <5, graded G2 if the total score is in the range of 6–9, and graded G3 if the total score is >10. In all the discussed cytological schemes, the cancers are graded G1, G2 and G3 to indicate the low (well-differentiated cells), intermediate (moderately differentiated cells) and high (poorly differentiated cells) malignancy tumors, respectively.

Robinson's scheme [36] consists of six cytological criteria namely cell dissociation, nuclear size, cell uniformity, nucleolus, nuclear margin and nuclear chromatin that were used to grade the tumors. A score of 1-3 is given to each of these criteria and the tumor is graded by summing up the scores. Consequently, the cancers that score in the range of 6 to 11 are graded G1, a score of 12 to 14 is graded G2, and a score between 15 and 18 is graded G3.

Howell's system [37] is similar to the BR hist-grading system but with a modification on the mitotic count criterion. In Howell's system, cancers are cytologically graded using three criteria which are: Tubule formation, nuclear pleomorphism, and MCC. With respect to the modification of mitosis count criterion, a score of 1 is assigned to cancers when there is 0-1 mitosis per 10 hpf, a score of 2 in case there are 2–4 mitoses per 10 hpf, and score 3 if there is >5 mitoses per 10 hpf. The final cancer grades are set to G1, G2, and G3 for scores in the range of 3–5, 6-7 and 8-9, respectively.

Six malignancy criteria, namely cellular pleomorphism, nuclear size, nuclear margins, nucleoli, naked tumor nuclei, and mitotic count were used to grade the tumors in Khan *et al.*'s system [38]. Here, a score of 1-3 is given to each of these criteria and the cancers are been graded G1 if the combined score is in the 6-10 range, grade G2 for a score between 11 and 14 and grade G3 if the score exceeds 14.

Finally, Taniguchi *et al.* [39]'s system consists of seven malignancy criteria, namely necrosis, cellular size, NCR, nuclear pleomorphism, nucleoli, chromatin density, and chromatin granularity. Likewise, all the criteria are scored in the range of 1-3 except the cellular necrosis which is scored 1 or 0 (for more details about the necrosis criterion see section 3.1.1). Here, grade G1 is assigned for the cancers if the total score ranges between 6 and 9, grade G2 if the total score is between 10 and 11, and grade G3 if the total score is grater than 11.

Despite the different opinions related to the extent of information that can be drawn from breast FNA biopsy, most cytopathologists agree that NG should be done on the FNA slide of primary and metastatic breast cancer [40]. This is because it provides valuable information to oncologists which allows them to better plan the treatment of the patients [23]. The concept of NG was introduced by Black *et al.* [42] in 1955. Fisher *et al.* [43] performed some subtle modifications on the NG and applied it in cytological smears. In the last few decades, numerous studies assessing NG in breast carcinoma have emerged. The cytological criteria of NG include architecture, background, dissociation of cells and many more, along with nuclear characteristics [40].

Based on the final score, the tumor is assigned a grade, menly G1, G2 or G3 indicating low (well-differentiated cells), intermediate (moderately-differentiated cells) and high (poorly-differentiated cells) malignancy, respectively.

Although FNA slides have been efficiently used for years for the diagnosis of breast lesions [4], the manual examination is a challenging task which necessitates a great deal of work from the pathologists and can lead to misclassified grades. Therefore, there is a need for a mechanism to verify the accuracy of the grade provided by the FNA manual exam which has led to increasing interest in computer-aided cytological image analysis. An automatic, objective malignancy diagnosis can assist

Cytological Grading Schemes	Cytological criteria
Fisher's Modification of Black's NG Scheme [34]	Anisonucleosis, nuclear membrane, nuclear chromasia, nucleoli, chromatin pattern and mitosis count criteria
Mouriquand's Grading Scheme [35]	Cellular characteristics, nuclear features, nucleolus and mitosis count criteria
Robinson's Grading Scheme [36]	Cell dissociation, nuclear size and uniformity, appearance of nucleoli, nuclear margin and nuclear chromatin criteria
MSBR Grading Scheme After Howell's Modification in mitosis count criteria [37]	Tubule formation, nuclear pleomorphism, and mitosis count criteria
Taniguchi et al. Grading Scheme [39]	Cell necrosis, cellular size, nuclear-cytoplasm ratio, nuclear pleomorphism, nucleoli, chromatin granularity criteria
Khan et al. Grading Scheme [38]	Cellular pleomorphism, nuclear size, nuclear margins, nucleoli, naked tumor cell nuclei and mitosis count criteria

Table 1: Cytological grading system examples [40].

inexperienced, overworked, or fatigued pathologists to avoid grading errors by providing a second expert opinion, particularly for uncertain cases that would require further manual examination by a pathologist. This can contribute to avoiding over-treatment for low malignancy grade patients. It can also allow for screening on a large scale to help identify systematic errors.

2.2 Computer-aided grading systems for breast cancer

Due to the alarming number of yearly breast cancer-related deaths of middle-aged women, computerized breast cancer diagnosis and grading have become one of the major research areas for both medical image classification scientists and clinicians. Therefore, and given the aforementioned shortcomings of manual examinations, automating diagnosis and grading of cancer cells is of prime

Cytological Grading Schemes	**Criteria scoring**	**Grade I per TS**	**Grade II per TS**	**Grade II per TS**
Fisher's *et al.* [34]	NaN	NG well differentiated nuclei	NG moderately differentiated nuclei	NG poorly differentiated nuclei
Mouriquand's [35]	0-3	<5	6-9	>10
Robinson's [36]	1-3	6-11	12-14	15-18
Howell's [37]	1-3 MCC: 1-Sc /(0-1MCC) 2-Sc /(2-4MCC) 3-Sc /(>5MCC) using 10 HPF	3-5	6-7	8-9
Taniguchi *et al.* [39]	1-3 except CN criterion NG 1 or 0	NG 6-9	NG 10-11	NG 12-19
Khan *et al.* [38]	1-3	NG 6-10	NG 11-14	NG 15-18

Table 2: Cytological grading system scoring methods [40, 41]. Abbreviations: TS: Total Score, Sc: Score, NaN:Not a Number, CN: 1 presence or 0 absence.

importance.

Researchers have been working on histology and cytology images of breast cancer for years. The analysis of histopathology and cytopathology images can usually be done using similar computer techniques to extract and classify image features in order to support the pathologists to achieve more accurate diagnosis results. This thesis focuses on cytological images of FNA biopsies of breast tumors. Image classification requires precise image analysis task to extract meaningful information from images. This analysis is mainly based on the concepts of image processing, pattern recognition, and machine learning algorithms. Obtaining a second opinion about the tumor diagnosis and grading has been facilitated by the classification task. Two fundamental approaches can be applied to accomplish the tumor classification task, namely the traditional and deep learning (DL) approaches.

2.2.1 Classical computer-aided grading systems

Automated breast tumors grading (tumor grade assignment) based on the traditional approach requires precise analysis of the cytological image content with the aim of extracting meaningful handcrafted features which will be used later for classification purposes [44]. Computer-aided cytological grading systems based on traditional methods make use of image processing and traditional machine learning algorithms to make a thorough analysis of the images of the tumors. Preprocessing, segmentation, feature extraction, feature selection, and classification are the basic components commonly used in automatic diagnosis or grading systems based on handcrafted feature engineering.

Image preprocessing :

Image Preprocessing is a fundamental step in every computer-aided grading system. The purpose of the preprocessing stage is to intelligently limit the abnormality search, without interfering with the crucial details of the image. Low-quality images might have severe noise or low-intensity contrast with weak edges. In order to identify the tumor region precisely, it is important to enhance the quality of the images by performing some image preprocessing techniques [45]. These techniques include image denoising using specific filters to reduce the image noise depending on the features of the image, image resizing to modify the range of intensity values of the image pixels, contrast image enhancement including histogram stretching to increase the contrast and equalization to enhance the contrast [46]. Image preprocessing is a very important step to aid precise image segmentation, especially for FNA images where they produced via different microscopic, magnifications as well as they composed different medical staining (see Figure 5). Furthermore, the effectiveness of the feature extraction task depends on the effectiveness of the segmentation results and in turn, this stage is the key to achieving accurate segmentation results

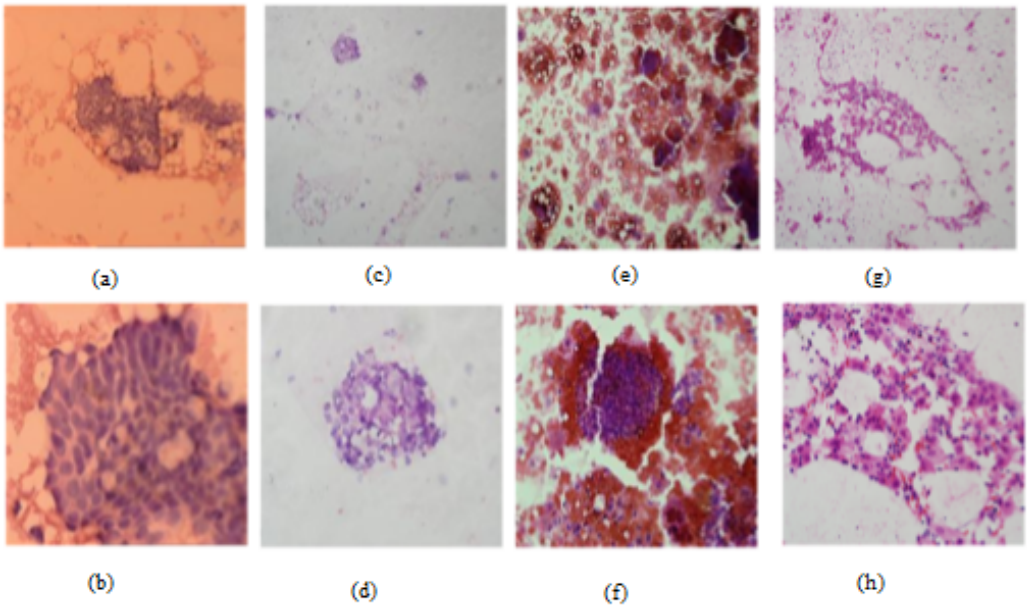


Figure 5: Different examples of the cytological images with both magnifications, (a),(c), (e) and (g) low magnification (100x power) and (b),(d), (f) and (h) high magnification (400x power) belong to G2 cases from JELEN_MERGE02 dataset .

Image segmentation :

In order to further identify and classify the disease, extracting proper cell nuclei regions from the images is mandatory [26, 47]. Image segmentation is a process of extracting significant information from the image background. Automated image segmentation is generally challenging due to the large variability (different microscopes, stains, cell types, cell densities) and complexity of the data (possibly time-lapse images, acquired at multiple wavelengths, using multiple microscopes, and containing large numbers of cells) as shown in Figure **??**. Therefore, the development of a generally applicable image segmentation method becomes a potential challenge [48]. Segmentation can be divided into two categories based on the image magnification, namely segmentation for low magnification images (LMI) and segmentation for high magnification images (HMI).

Image segmentation for the images with low magnification uses a group-based feature extraction task and is not capable of performing in-depth segmentation of the image objects. On the other hand, highly magnified images require more sophisticated methods of segmentation as they are used for extracting cell nuclei related features. These features not only require accurate nuclear boundary estimation (to calculate shape-based features) but also staining intensity level determination of all nuclei as well as the extraction of other features such as cell cytoplasm features. Thus, accurate nuclei segmentation results yield an accurate estimation of different types of cellular and nuclear features.

Common segmentation approaches which have been used on cytological images in early segmentation approaches include intensity thresholding, edge detection and region growing [10, 48, 26]. Intensity thresholding ensures that cells have distinct intensities compared from the background. It is carried out by using threshold values to separate objects from the background.

The global thresholding technique of Otsu's method also can be used. This method uses the gray level histogram to find the optimal threshold value that can be used to minimizes the weighted within-class variance or maximize the weighted between-class variance of the thresholded foreground and background pixels [49, 47]. However, this method has some drawbacks such as it partitions the grayscale histogram of an image into two classes, which are not equivalent to real environmental problems where images usually have a different number of classes. Further, it requires uniform illumination that is difficult to achieve on medical images, but it can be used as an addition to any other method. In other cases, edge detection (first-order differential filtering) is used. In general, neighbor pixel intensity analysis or gradient variations are applied to identify the boundaries between objects and background. Most common representations of the neighborhood of the pixels are the four-connected neighborhood and the eight-connected neighborhood. Edge detection usually uses special filters such as Canny and Sobel [26]. A linking procedure [48] follows thereafter. The main objective of edge linking is to shape significant borders. Elimination of the weak edges and linking broken edge segments are then carried out [26].

Hough transform (HT), allowing detection of image shapes, is another edge-based segmentation approach. HT (mostly circular and elliptical) performs well when the object's edges are not fully preserved [10, 50, 51], as well as when the images are noisy in nature. Circular and Elliptical HT differ in terms of computational complexities (Five parameters for Elliptical against three for Circular). Both have been chosen based on the assumption of nuclei shape being close to either elliptical or circular [52]. Depending on the analytical description of the shape of interest, HT works on the principle of mapping image points to a Hough space (accumulator). A limitation of this approach is that the accumulator space becomes larger as the number of free parameters increases. According to our experimental preliminary results, this algorithm performed poorly and failed for 70% of the images in the FNA image.

Machine learning based algorithms such as clustering-based and active contour techniques have been recently proposed, with the aim of improving the segmentation results. Logistic regression segmentation algorithms can be applied when prior knowledge of objects such as training samples or the number of clusters is available. Unsupervised learning such as k-means and mean shift clustering can be applied to group image points to different objects, without a set of labeled samples. K-means clustering is one of the most common clustering algorithms used to carry out segmentation. This

approach involves grouping a cluster of image points into k clusters by using the following objective function minimization:

$$\sum_{k=1}^{K} \sum_{i \epsilon S_k} | I_i - \mu_k |^2 \tag{1}$$

Where I_i is the intensity of the image point X_i in the class S_k, μ_k is the current mean of S_k, and $P_k = \sum_{i \epsilon S_k} p_l$, where p_l is the normalized histogram of unit intensity.

On the other hand, mean shift clustering does not assume prior knowledge of the number of clusters. The image space can be characterized by a certain probability density function while carrying out the segmentation of the 2-dimensional red green blue (RGB) image. All the pixels in the image can be represented as points in a 3D grid. The probability density function (PDF) runs over each point in the grid with a speechified bandwidth. This results in a PDF surface with specified peaks. The mean shift algorithm will try to group all the points of the image close to their respective peaks, thereby making clusters of data. This resulting cluster, when mapped back to the image, gives the segmented result.

The kernel density estimator for n points (X_i,i=1,2,...,n) is defined as:

$$f(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K(\frac{X - X_i}{h}) \tag{2}$$

Where K(X) is the kernel with bandwidth h. the gradient ascent procedure is guaranteed to converge to a point where $\bigtriangledown f(X) = 0$, i.e., the local maximum (mode).

In addition to the segmentation techniques described above, another popular approach used for specifically reflecting the boundaries of input objects is active contour or snake. By using this approach the segmentation can be done with high accuracy and provide closed and smooth contours [26]. Once the initial boundary contours are fed as input to the algorithm, the contours are mapped to the desired original boundaries of the image objects. The fitting process of the active contours specifically takes into account two types of energy: Internal energy E_{int} defined within the curve, and External energy E_{ext} calculated from the image. E_{int} maintains the curve smoothness (during deformation), whereas E_{ext} measures the edginess of the surface through which the image contour passes. The purpose is to minimize the total energy, which is obtained by adding the two associated energies as follows[10, 53].

$$E_{sum} = E_{int} + E_{ext} \tag{3}$$

Gray level quantization is a segmentation approach which relies on the nuclei textural description. In order to generate the gray level co-occurrence matrix (GLCM) to estimate texture features [10, 54],

second order statistics are used. Considering pairwise combinations of grey levels, the conditional j thatoint probabilities C_{ij} are calculated (see 4) for the image spatial windows. The schema assumes that the inter-pixel distance is known. The C_{ij} for the image spatial windows are defined as

$$C_{ij} = \frac{P_{ij}}{\sum_{i,j=0}^{G-1} P_{ij}} \tag{4}$$

Where P_{ij} is the frequency of occurrence of the two grey levels i, j and G is the number of quantized gray levels. A co-occurrence dependency matrix is generated, with the element (i, j) of the matrix representing the probability C_{ij}. From the dependency matrix, some features, such as the entropy, contrast, correlation and energy, can be extracted with the purpose of identifying textures.

After the completion of the segmentation phase, a set of different features can be extracted from the detected cell nuclei regions as well as some other cell organisms such as the cytoplasm regions. The features can then be introduced to classifiers as an input vector (Classification Stage). The extracted features should correspond to the cytological-based characteristics as mentioned earlier (see Table 1) in the Introduction chapter. To achieve this goal, the selected features should reflect the biological changes (morphological, polymorphic and textural) of cancerous cells, which closely reflect pathological cytological characteristics used by the CGSs.

Features extraction :

Feature extraction is a process through which a set of new features is created from the original image. This process is mandatory for general image processing and computer vision applications such as object detection, pattern recognition, and image classification. For the malignancy grading of the breast cancer problem, as mentioned earlier, usually different magnifications of images are used during the feature extraction process (100x,200x,400x and 600x). However, in this thesis, due to the collected dataset, we considered only two magnifications of images, namely LMI with 100x power and HMI with 400x power. Among the numerous features that can be extracted from cytological or histological images of breast cancer, binary features (area, perimeter, convexity, eccentricity, centroid, orientation, projection, etc.), momentum-based features (features that are rotation scaling translation (RST) invariant), histogram-based features (describe the occurrence frequency of intensity values in the image), textural features (to measure the texture information of the nuclei) that can be obtained using different methods such as second-order statistical using GLCM and GLRLM (describe a combination of repeated patterns of pair of pixels with a regular frequency), local binary patterns (determines a label for each pixel in an image based on estimating a threshold value of the 3x3 neighborhood of each pixel with respect to the center pixel and uses the histogram of the labeled pixels as a texture descriptor), and wavelet coefficient (low or high-frequency features

extraction from the wavelet decomposition of an image), are the most important ones [10, 26, 55, 56].

Features selection and dimensionality reduction :

To determine the most discriminative features of the initial set of candidate features, feature selection or dimensionality reduction (DR) stages are important to improve the efficiency of the classification system [26]. Feature selection is different from DR since feature selection methods include and exclude attributes present in the data without changing them, whereas the DR method creates new combinations of attributes [57]. Feature selection for supervised classification tasks can be accomplished on the basis of the correlation between features. Such a feature selection process can be beneficial to a variety of common machine learning algorithms [58].

Algorithms like fast correlation-based feature selection (FCBF), Markov blanket filter (MBF), and correlation-based feature selection (CFS) use the discriminating criteria for feature selection. Correlation coefficient or statistical tests like t-test or f-test are used in these approaches to filter the features. These methods are very simple, fast, and also independent of the classification algorithm. Some of the other significant approaches such as sequential forward selection (SFS) where relevant features are added one by one, beginning with an empty set, sequential backward elimination (SBE) where irrelevant features are removed in backward direction and genetic algorithms which are optimization algorithms starting from a group of points coded as a finite length alphabet [59].

In order to transform feature points to a low dimension, DR tools are commonly used. The feature points are transformed in order to yield a feasible selection and classification (linear and non-linear techniques). The linearly separable points in the feature space, linear techniques like principal components analysis (PCA), linear discriminant analysis (LDA), and MDS are used. On the other hand, for inherently nonlinear biomedical structures, nonlinear techniques like spectral clustering, isometric mapping, locally linear embedding (LLE), and Laplacian eigen maps (LEM) perform best. In order to identify the malignancy level of a tumor, classification algorithms can be applied based on the simplified feature vectors obtained by DR techniques. The classification process begins after a set of good discriminative features are selected. The purpose of this phase is to identify or classify the grade of malignant breast tumors.

Classification task :

In the classification stage, the classification algorithms compare a set of pre-derived training sample features to the input image features [26]. The main goal of a classification method is to identify the class to which a certain case belongs. In other words, an FNA cytological image (the used dataset in this thesis) is classified using classifiers that take a feature vector as an input and come

up with the specific grade of malignancy that belongs to a certain case. Pattern classification is the approach used to discriminate various classes of patterns. Each classification method is expected to misclassify some of the input patterns. Different metrics are used to measure the performance of classification algorithms such as the Error Rate (the lower the error the better the classification obtained [10]), accuracy (the number of correctly predicted samples divided by the total number of predicted samples), ROC curve (is a probability curve that summarizes the trade-off between the true positive and false-positive rates for a classification model), etc. In the classification process, the dataset is divided into three subsets (training, validation, and testing) for the purpose of examining the accuracy of the classification algorithm. The classifiers learn the optimal parameter (e.g. weights of connections between neurons in an ANN) settings while using the training subset. The validation subset is used to validate and select the best model (that minimizes overfitting which occurs when a classification model doesn't generalize well from training dataset to unseen dataset) for the classification task. Finally, the testing subset is used to report the algorithm prediction errors.

To classify the data (feature vector) there is a wide selection of classifiers that can be used. To start, the naive Bayes (NB) classifier [60] is one of the most common techniques that consider the prior and conditional probabilities associated with the dataset [10]. It is based on the assumption that all the relevant probabilities are known and that the attributes of a feature vector are statistically independent [61].

The K-nearest neighbors (KNN) [62] is another popular classification algorithm that relies on the closest association between the pattern (closest class, the distance among k-neighbors) and the neighbors based on the Euclidean distance calculation. Here, the Euclidean distance is used to calculate the distances between the training samples and the test samples [10, 63]. The class is assigned to the sample for which the distance is smallest.

The decision tree (DT) algorithm [62] is a classifier where the top node in a decision tree is known as the root node, interior nodes represent the features and branch represents a decision rule. Each leaf node represents the outcome (contains a response such as 'true' or 'false). The decision path is the path from the root to the correct leaf.

Individual decision trees tend to overfit the dataset. However, the random forest of decision trees (RFDT) [64] can be used instead to reduce the effect of overfitting and improve generalization. RFDT is in fact an ensemble of DTs forming one of the most used supervised learning algorithm because of its simplicity and flexibility. Usually, this algorithm is trained with the Bootstrap aggregating (bagging) method which used to improve the stability and accuracy of algorithms by combining

weak models to increase the overall result. To split a node, only a random subset of the features is usually considered by the algorithm.

Moreover, the artificial neural network (ANN) [60] is another significant classification method that should be considered in this work. Depending on the architecture, neural networks can be classified into single-layer neural networks (one layer of neurons), where each neuron connects directly to an input variable and contributes to an output variable, multilayer networks, also called as a multilayer perceptron which is a cascade of single-layer perceptrons (at least two separate layers and output signal) and Recurrence neural networks (the signal depends on the output through the loopback). Two significant hyperparameters are the number of layers and the number of nodes in each hidden layer controls the architecture of an ANN. These hyperparameters experimentally determine according to the underlying problem and the dataset. For example, for a linearly separable dataset (a line can be used to separate two classes), a single-layer neural network can be used to classify samples. Whereas, for the not linearly separable dataset which the case for most real-world problems, a multilayer perceptron can be used to classify the samples of different classes by drawing shapes around samples in some high-dimensional space to classify these samples. A traditional feed-forward multilayer neural network consists of an input layer (receives input variables), an output layer (produces the output variables), and many hidden layers (layers of neurons between the input and output layers). An ANN works by combining a few neurons together in such a way that the neurons can interact with one another, making it capable of processing input data and providing a decision (output) [10, 65, 66]. An example of a standard multilayer ANN is shown in Fig. 6.

In order to separate two or more classes of patterns or data points, support vector machines (SVM) [48, 68] can be used. It is another discriminative method defined by a separation of the two classes through hyperplane. The algorithm works under the hypothesis that finding a better hyperplane (maximizes the margin) that gives the largest distance to the training examples. The boundary points are called support vectors. Nonlinear feature space with the kernel Trick can be visualized as an extension of the original linear SVM. Kernel functions are based on inner products of two feature vectors [26].

2.2.2 Deep learning-based approach for computer-aided grading system

Deep learning has achieved high accuracy in different applications including in medical image classification tasks. According to numerous studies related to DL networks, CNN or ConvNet architectures are better for image classification tasks since they have some additional layers besides those in the standard neural network (input, hidden and output layers). Those additional layers use spatial structures of the input image to extract image features which are used later for image classification

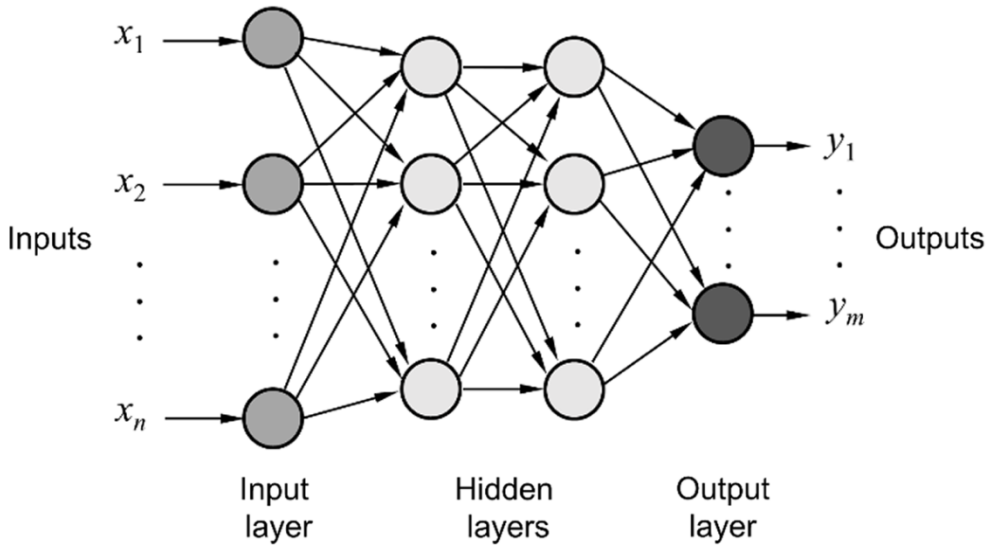


Figure 6: Example of typical structure of a feed-forward multilayer neural network composed of input, output and two hidden layers taken from [67].

purposes. Hence, different CNN architectures can be used to build a malignancy grading system to classify our FNA biopsy images. Using a CNN architecture, a classification model can learn to perform classification tasks directly from images without the need for the manual feature extraction procedure. According to the classification problem and the amount of training dataset, we can construct a network from scratch and train it to classify the FNA images or we can use a pre-trained model to perform transfer learning and retrain the model to classify our FNA image dataset.

CNN algorithms can have tens or hundreds of layers with a huge number of learnable parameters. Each layer learns to detect different features of an image. For each image, different windows or filters of different resolutions can be utilized to extract various low-level (colors and edges) and high-level (more specific to image objects) features according to the selected windows. The output of each convolved image is used as an input to the next layer.

Although there are different types of CNN architectures, they share common significant layers, namely the convolutional layer which extracts features from the input image by convolving it with a set of filters based on performing a wise multiplication (multiplying two matrices element by element), the ReLU which maps negative values to zero and preserves positive values to speed up the training process (uses to introduce nonlinearity to a classification model after performing linear operations during the convolutional layers), the pooling layer that reduces the spatial size of the feature maps independently to reduce the amount of learnable parameters and computations, and

the fully connected layer which computes the class scores based on the combined feature maps. Further, there are three important hyperparameters that control the size of the output volume for any CNN model. These hyperparameters are the depth size which corresponds to the number of filters required in each layer, the stride that represents the amount of sliding or shifting of the filters over the input volume, and zero-padding by means of adding zero around the border of an input volume to preserve the original size of the input volume.

Regarding the computer-aided grading system in this thesis, different ConvNets can be used to determine the malignancy level of FNA biopsies of breast cancer. However, for small training sets, it is not effective to train a network from scratch (high computational cost due to a huge number of parameters) to get high accuracy for the specific problem at hand. Therefore, it is more efficient to perform a transfer learning with pre-trained deep network models such as AlexNet, GoogLeNet (Inception-v1,v2,v3 or v4), Residual Networks (ResNet18, 50, 101), and VGG-19, to quickly learn the new classification task using our dataset. Transfer learning help in building an accurate model in a very timesaving way by starting the learning process based on considering low-level patterns that have been learned previously via solving a different problem using pre-trained CNN models. To begin with, AlexNet (2012) [69] is 8 layers deep, 5 of them are convolutional layers, which could be followed by max-pooling layers, and 3 fully-connected layers. Further, it is composed of 60 million parameters and 650,000 neurons (see Fig. 7). Furthermore, it has been trained on more than a million images from the ImageNet database and can classify images into 1000 object categories. The network considered outperformed the previous state-of-the-art in the imageNet large scale visual recognition competition (ILSVRC)-2012 since it has achieved the lowest top-1 and top-5 error rates of 37.5% and 17.0%, respectively.

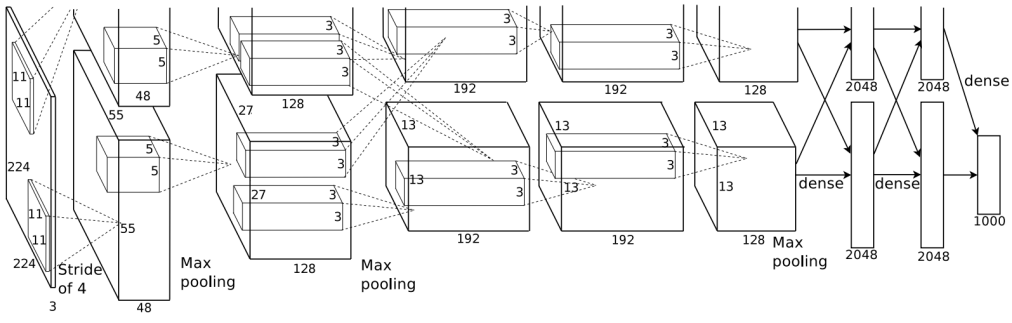


Figure 7: The overall architecture of AlexNet CNN taken from [69].

Another popular and better speed and accuracy performance image classification CNN is GoogLeNet (2014) [70], that has been trained on 1.2 million images from the ImageNet database and can classify images into 1000 object categories. GoogLeNet was created based on the idea that some

of the activation maps (the output of a given convolution filter) are unnecessary (with zero value) or redundant due to the high correlations among these activation maps. In most deep learning architectures, this causes a weak connection among the mentioned activation maps which can be interpreted in such a way that not all the activation maps will have a connection with all the inputs. In this regards, inspection models of GoogLeNet handle this issue by approximating this weak connection using a spatial organization of different sizes of convolution layers to capture different details at varied scales (5×5, 3×3, 1×1). In other words, different filters in different layers are trying to activate different parts of an image. Some filters seeking on edge detection, others are detecting a particular region of the image such as its central point and others are detecting the image background. Thus, rather than manually determining which type of convolution is the best to use at each layer, the developers included all types of convolutions and let the model decide which type is best for a particular convolution layer.

Several versions of the GoogLeNet network have been constructed. The original version of GoogLeNet, the so-called Inception-v1, consists of 22 layers in depth and has 7.0 million parameters. The network was considered better than the previous state-of-the-art in the ILSVRC-2014 competition by achieving a top-5 error rate of 6.67% (see Fig. 8).

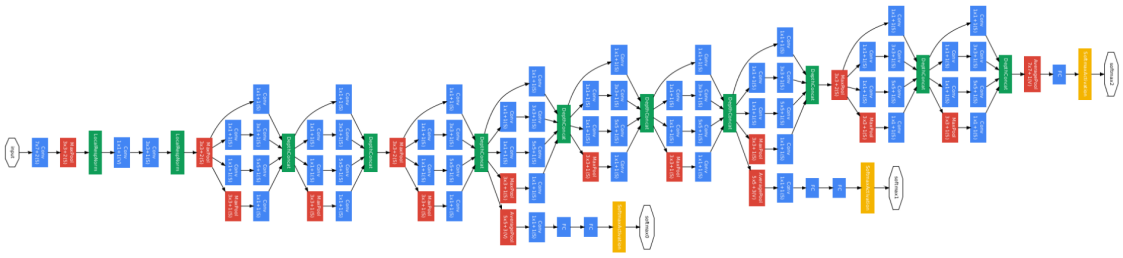


Figure 8: The overall architecture of GoogLeNet-Inception-v1 network taken from [70].

The Inception-v1 model of GoogleNet network consists of a spatial organization of convolution layers, different in sizes and types from convolution layers in other networks such as AlexNet and ResNet, called Inception model where different filters are utilized in parallel. Precisely, each Inception consists of 1×1 convolution, 3×3 convolution, 5×5 convolution, and 3×3 max pooling. These layers work together for the previous input and again stack together at the output (see diagram (a) of Fig. 9). However, in GoogLeNet or Inception-v1, 1×1 convolution is used as a dimensionality reduction for feature maps (decrease the number of learnable parameters by reducing the number of computations in the network) to reduce the input to large convolutions such as (3×3, 5×5). Thus, to reduce the dimensionality in the originally introduced Inception model, 1×1 convolution was added to the

Inception model which aimed to decrease the number of parameters in each Inception (see diagram (b) of Fig. 9). Further, by using local response normalization and a global average pooling at the end of the network, the top-1 accuracy of GoogleNet has been improved by about 0.6%.

For Inception-v2 and Inception-v3, the authors successfully increased the accuracy and reduced the computational complexity in these two versions by reducing the dimensions and adding some techniques such as batch normalization and factorization. To begin with, the Inception-v2 model [71, 72], also known as the Inception-batch normalization (BN) module, was developed with the aim of reducing many of the calculation operations and also to avoid the overfitting problem that can occur in the Inception-v1 model. Basically, in CNN models, layer's input is influenced by parameters (the weights and biases of the neurons) are shared in all the input layers. This leads to an unstable distribution to internal layer inputs (due to weights update) that create the so-called internal covariate shift [73]. Thus, BN helps in speeding up the training process of these models by normalizing the inputs of each layer to alleviate this Internal Covariate Shift problem. More precisely, BN provides some regularization effect which reduces generalization error that in turn minimizes overfitting effect. Thus, in this version of GoogLeNet, firstly, the BN technique was introduced and a 5x5 convolution was factorized into two 3x3 convolution operations to improve computational speed as shown in Fig. 10. This step reduced the computational cost in this version as compared to Inception-v1 where a 5x5 convolution is 2.78 times more expensive than a 3x3 convolution according to the study. Secondly, to further reduce the dimensionality and avoid overfitting problems, an asymmetric convolution was performed by applying additional factorization process to factorize large convolutions into small convolutions. Thus, $n \times n$ factorize into $1 \times n$ and $n \times 1$ with the aim of decreasing the computational cost (see Fig. 11). As an example, consider that n=3 to achieve the equivalent of the previous diagram (b) of Fig. 10. Thirdly, the filter banks in this module were expanded wider to remove the representational bottleneck of the Inception-v1 style as shown in Fig. 12.

The Inception-v3 model [72] consists of 48 layers. In this third version of GoogLeNet, all of the three upgrades steps that mentioned in Inception-v2 model were maintained as well as the following upgrades were added in this version: RMSProp optimizer (to find the optimum values or the best solution), factorized 7x7 convolutions into three 3×3 convolutions (to reduce the dimensionality), BN-auxiliary instead of BN has been introduced which allows the normalization of the fully connected layer not only the convolution layers like in previous versions of GoogLeNet and label smoothing (regularizing term) was added to the loss function to prevent the network from overfitting.

The Inception-v4 model [74] is the most simplified and uniform architecture of the GoogleNet network that introduced the reduction blocks concept. A reduction blocks idea was introduced in this version

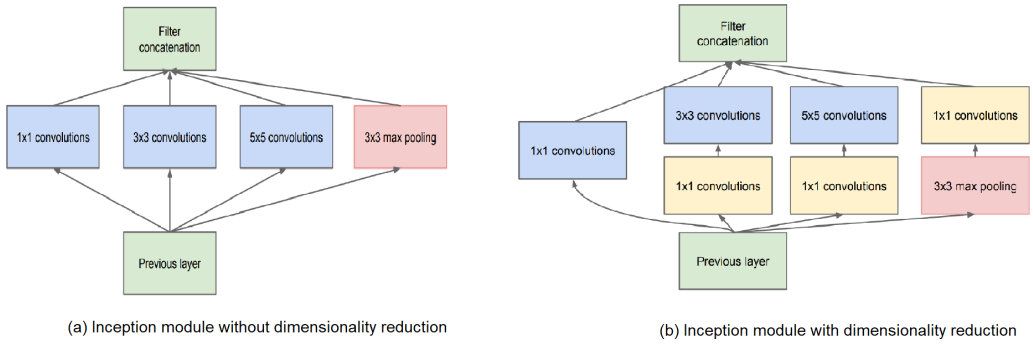


Figure 9: Original and modified Inception style for GoogLeNet-Inception-v1 model, (a) Original Inception without dimensionality reduction (b) Modified inception with dimensionality reduction taken from [70].

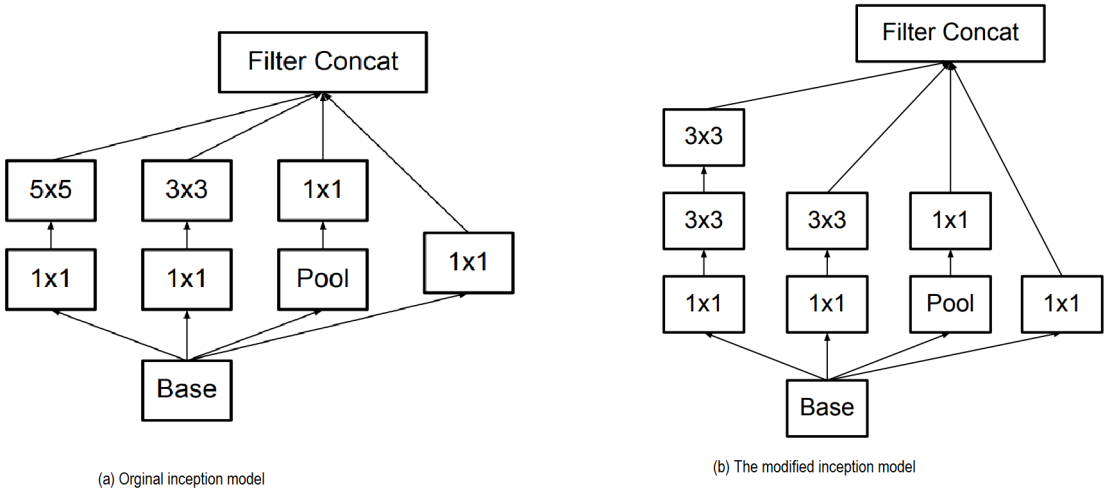


Figure 10: The Inception-v2 model or Inception-BN network of GoogLeNet, (a) The Original Inception module and (b) The modified Inception module with dimensionality reduction where the 5×5 convolution in diagram (a) is replaced (factorized) by two 3×3 convolution in diagram (b), taken from [72].

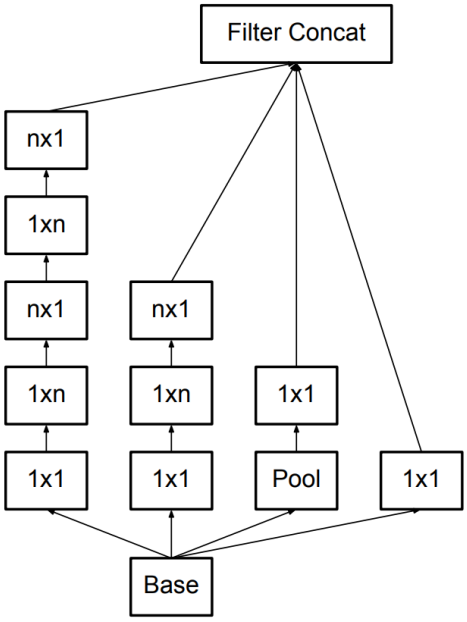


Figure 11: The GoogLeNet-Inception-v2 after factorizing each $n \times n$ convolutions (consider n=3 to achieve the equivalent of the diagram (b) of Fig. 10) into $1 \times n$ and $n \times 1$ in this diagram, taken from [72].

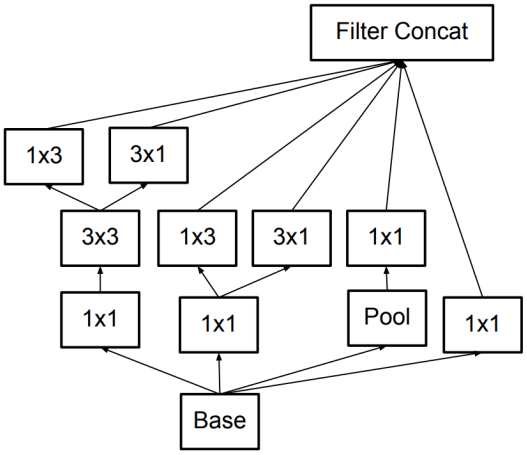


Figure 12: The GoogLeNet-Inception-v2 after making the Inception module wider by expanding the filter bank outputs, taken from [72].

where the large scale structure of the Inception-v4 networks was shown in diagram (a) of Fig. 14. Using this reduction idea the authors explicitly modified the width and height of the grid which was not possible in the old versions. This version has three main Inception modules, named Inception-A (35x53 grid decreased to 17x17 dimension grid in Reduction-A block), Inception-B (17x17 grid decreased to 8x8 dimension grid in Reduction-B block) and Inception-C (see Fig. 13). Both the original blocks of Inception A and B as well as their reduction versions are shown in diagram (a) of Fig. 14 and diagram (b) of Fig. 14 illustrates an example of the Inception-A block that was used of the inception-v4 network as reported in [74].

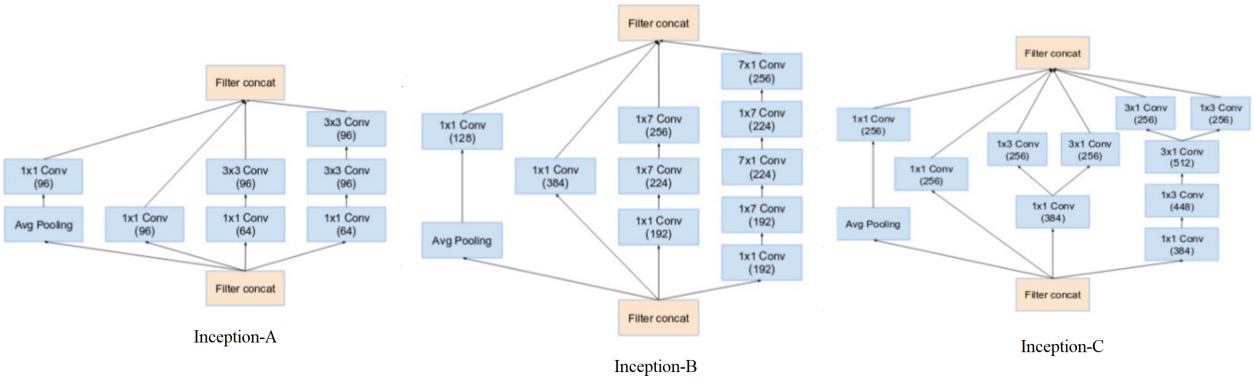
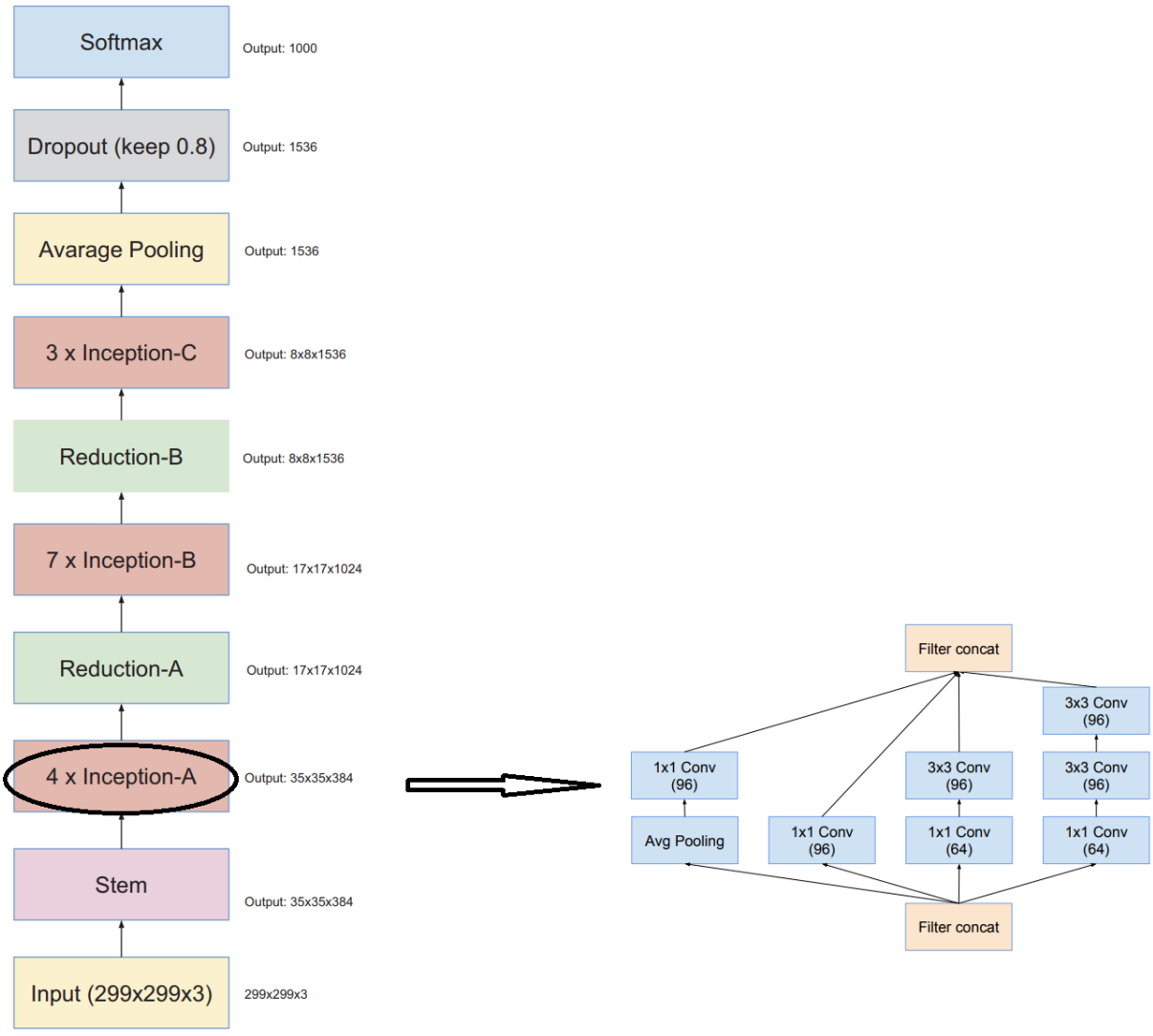


Figure 13: The Inception modules A,B and C used in Inception v4 taken from [74].

ResNet is another significant deep learning method. The idea behind ResNet, shown in Fig. 15, is to introduce an identity shortcut connection that aims to skip one or more layers due to the vanishing gradients problem associated with very deep neural networks. The vanishing gradients problem refers to the difficulty occur during the network training and convergence of CNN models where the gradient will be vanishingly small which in turn prevent the parameters update of the early layer required during training these models. Thus, ResNet can efficiently train networks with 100-1000 layers and still achieve reliable performance. Thus, ResNets are built based on the concept of dividing a very deep neural network into small blocks of networks linked through skip or shortcut connections to form a bigger or deeper network (see Fig. 16). The network has been trained to extract rich feature representations using a subset of 1.2 million images of ImageNet and can classify the images into 1000 object categories. Further, the network was validated using 50,000 validation images and tested using 100,000 testing images. According to the network depth, there are five versions of residual networks that are ResNet18 (18 layers deep), ResNet34 (34 layers deep), ResNet50 (50 layers deep), ResNet101 (101 layers deep), and ResNet152 (152 layers deep) as reported in [75]. Using an ensemble model, ResNet won 1st place in the ILSVRC-2015 classification competition with the top-5 error rate of 3.57%. According to the ResNet developers, the ResNet-101 network has



(a) The overall of the Inception-v4 network

(b) The Inception-A block

Figure 14: Inception-v4 network, (a) The overall diagram of the Inception-v4 network, (b) The Inception-A where used in this Inception-v4 verion of GoogLeNet taken from [74].

fewer parameters than other deep networks and it can converge well. Further, it achieved the best classification error with 6.43%.

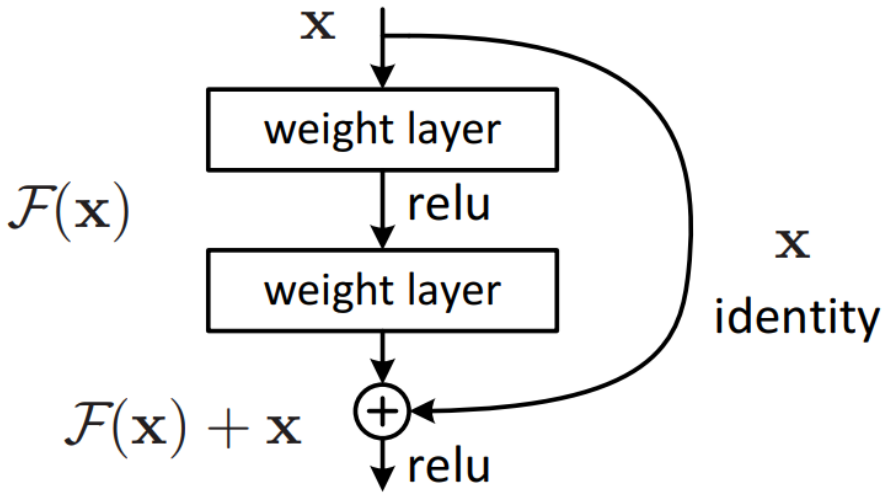


Figure 15: A building block of Residual learning taken from [75].

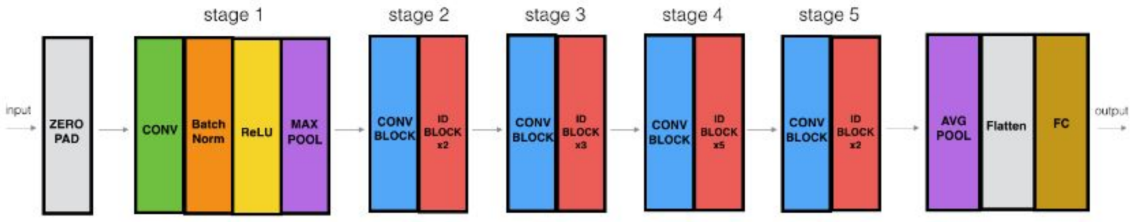


Figure 16: The overall architecture of ResNet-50 taken from [76].

2.3 Performance evaluation for binary classification problem

Many standard performance measures can be used to evaluate classification systems. Thus, with the aim of evaluating the performance accuracy of different classification systems, whether they are created based on traditional or deep learning methods, various techniques such as leave-one-out or general k-fold cross validation are used. Further, a confusion matrix can be computed to describe the performance of a classification model. More precisely, it is a table used to summarize how many samples of different classes were correctly and incorrectly classified by the model (see Table 3). True positive (TP), false positive (FP), true negative (TN) and FN are the four categories into which the grading results of the testing subset can be classified.

		Predicted labels		
		Negative	Positive	Total
Actual labels	Negative	TN	FP	$TN + FP$
	Positive	FN	TP	$FN + TP$
	Total	$TN + FN$	$FP + TP$	

Table 3: Example of confusion matrix table for two class problem .

To begin with, TP represents the number of positive class samples that are correctly classified as positive, FP denotes the number of negative class samples that incorrectly got classified as positive, TN is the number of negative class samples that were correctly classified as negative, while FN indicates the number of positive class samples that got incorrectly classified as negative [77]. The testing subset grading results can be classified into one of the four aforementioned categories. Based on these categories, different performance matrices, such as accuracy, specificity, and precision can be computed from the confusion matrix (see Fig. 17) [78]. Accuracy measures the proportion of all true samples to total samples. The sensitivity scales the system's accuracy in identifying positive samples, while the specificity grades the accuracy of the system in finding negative samples. The precision measures the proportion of true positive samples to all positive samples. Accurate computerized diagnosis is a reflection of high values of both sensitivity and specificity. The area under the receiver operating characteristic (ROC) curve is a result of the sensitivity and specificity combination where the larger the area, the better the performance of the system [78].

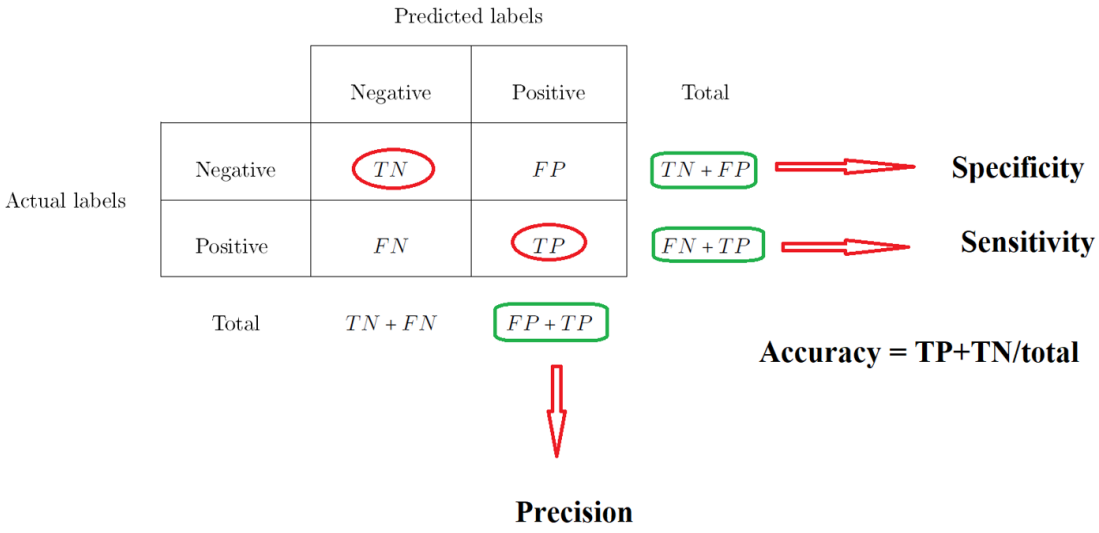


Figure 17: The four different performance metrics that be calculated using confusion matrix table

.

2.4 Literature survey of breast cancer diagnosis and malignancy grading problems

In this section, we address the related literature from three aspects, namely the traditional method for classification, unbalanced class distribution problems, and DL classification system.

With regards to pathology-based studies, many of these studies recommended that cytological grading should be part of all FNA reports of breast carcinoma so that preoperative prognostication could be evaluated [32, 33, 29]. However, the features for cytological grading are not well established. There have been different conclusions regarding the importance of these features and their correlation with the standard grading system on histology.

Robinson *et al.* [36] noticed that there was a direct correlation between all the cytological scores in their grading scheme and the histological scores of the BR scheme. They concluded that cell dissociation and appearance of nucleoli are the most powerful predictive factors. With reference to Robinson's schema, Saha *et al.* [40] found that, except the cell size and nucleoli, the cytological parameters of Robinson's system had significant importance in predicting the final cytological grade. Taniguchi *et al.* [39] found that CN had a non-significant or weaker correlation with the final cytological grade while Chhabra *et al.* [79] found a strong correlation between each cytological feature and cytological grade.

To this day, there is no reliable method for cytological grading which closely matches the histological grading system [32]. The clinicians often put little importance on this valuable prognostic parameter due to subjective evaluation and absence of uniformity. Pandey *et al.* [32] evaluated and compared Robinson's and Mouriquand's cytological grading systems for breast carcinoma and examined their correlations with Nottingham modification of Scarff Bloom Richardson (SBR) histological grading. The authors reported a concordance of approximately 83.33% between Robinson's system and SBR, and of 66.66% between Mouriquand's system and SBR. Further, they stated that Robinson's system showed a diagnostic accuracy of 90% with 91.30% sensitivity while Mouriquand's approach had an accuracy of 76.66% with 95.65% sensitivity. Das *et al.* [33] conducted a double-blind study to evaluate Robinson's and Mouriquand's cytological grading systems. The main objective of this study was to determine which one of these two systems was corresponding best with the histological BR-scheme. According to their results, Robinson's system showed a diagnostic accuracy of 80.76% with 77.77% specificity while Mouriquand's method gave 84.60% accuracy with 33.33% specificity. According to the discussed pathology-based studies, although the examined CGS systems were found to have similar concordance with histological grading, Robinson's method was considered the best because of its simplicity and reproducibility.

Saha *et al.* [40] evaluated and compared the six published cytological grading systems (see Fig. 1) with Nottingham modification of the SBR's method. The comparison was performed by testing concordance, association, and correlation with the result obtained by the SBR histological grading method using a dataset of 57 patients with breast carcinoma. The study reported that all six cytological grading systems correlated strongly and positively with the SBR grading method. Further, Robinson's system achieved the best correlation and concordance with 79.9% and 77.19%, respectively, followed by Mouriquand's method which obtained 71.5% and 77.19% correlation and concordance, respectively. The aim of reviewing pathology-based studied was to figure out which of these six systems is more correlated to the histological grading systems that recommended among pathology-based study to use as references to evaluate the computer-aided grading systems based on cytological images.

With regards to computer-aided grading systems for breast cancer, histological grading studies are first reviewed because a large focus of pathological image analysis has been on automated grading for histopathological slides. Moreover, these studies provide a more comprehensive view of the disease and its biological effect on tissues structures (the underlying tissue architecture being preserved by the preparation process) and can help acquire some ideas about the most effective techniques and algorithms that can be used as first attempts to examine the malignancy grading task on our FNA dataset. To begin with, Distinguishing method between graph-based features and high-grade for

breast cancer histology specimens, based on the presence or absence of lymphocytic infiltration, was proposed by Basavanhally *et al.* [80]. The authors automatically detected Lymphocytes through a segmentation scheme, comprised of NB classifier and template matching. The graph-based feature approach, in conjunction with an SVM classifier, was used in the classification stage and achieved a classification accuracy of 89.50%. Doyle *et al.* [81] presented an image-analysis approach to automatically distinguish low and high grades of breast cancer from histology images. The methodology involved the extraction of over 3,400 image features (textural and nuclear architecture based) from a database of 48 breast biopsy tissue images (30 cancerous and 18 benign). In order to reduce the dimensionality of the feature set, spectral clustering was used. In the classification stage, SVM was used for two purposes, namely to differentiate benign and malignant cases and to facilitate the categorization of cancer malignancy grades. An accuracy of 95.8% in distinguishing cancer from non-cancer using texture-based characteristics, and of 93.3% in distinguishing high from low grades of cancer using architectural features were achieved. Naik *et al.* [82] contributed to the automated detection and segmentation of nuclear and gland structures in histology image slides. Image information retrieved from three scales, namely low-level information (pixel values), high-level information (relationships between pixels for object detection), and domain-specific information (relationships between histological structures) were integrated. The NB classifier and level set algorithms were used to extract low-level information. The NB classifier determines a contour to search around the likelihood pixels that belong to an object of interest. Meanwhile, the high-level information was retrieved using level-set and template matching algorithms (shape models used to identify glands and nuclei from the low-level likelihood scenes). Further, different morphological and nuclear features were extracted for automated grading of various cancers like prostate and breast and for differentiating cancerous and non-cancerous breast histology specimens. The strength of the proposed model is that it incorporates low-level, high-level knowledge, and structural constraints imposed thorough domain knowledge. According to the study results, the highest accuracy was 95.19% for grade 3 vs. grade 4 classification task.

As highlighted earlier, the main objective of this thesis is to propose a computer-aided grading system for FNA biopsy of breast cancer to precisely determine the malignancy grade of cancer as early as possible. The existing literature is quite limited with regards to this problem. Cytological images often result from the least invasive biopsies. Additionally, the characteristics of cytological images, i.e. presence of isolated cells along with their clusters and absence of more complicated structures (tissue and glands, which are available in histological images) make it easier to analyze these specimens compared to histology. In the context of our problem (malignancy grading for FNA biopsy of breast cancer), we specifically analyze the work of Jeleń *et al.* [83]. Their work is the only relevant contribution to the problem stated above.

Jeleń *et al.* [83] presented the first automatic malignancy grading framework to identify the malignancy grade for breast cancer cytological images of FNA biopsies. Their approach was based on adapting the histological BR scheme [19] to grade cytological FNA slides. The BR scheme [19] grades breast carcinomas by adding the scores of three characteristics: Degree of Structural Differentiation (e.g. tubule formation), Nuclei Pleomorphism, and MCC. Based on the obtained value for the three added scores mentioned, the tumor is allocated one of the three grades: G1, G2, or G3 representing low, intermediate, and high malignancy, respectively. Histological grades describe features related to tumor differentiation and proliferation (from well-differentiated for a low grade to poorly differentiated for a high grade). From the results presented in the paper, the best average accuracy which is of 82.7% was achieved using a multilayer perceptron (MLP) classifier and the other obtained results from SVM, SOM and PCA classifiers were less by about 1-2% according to the study.

Recently, aiming to improve their previous work [83], Jeleń *et al.* [84] studied the features extracted from FNA biopsy images to determine their relative discriminatory power and cross-correlations with the objective of reducing the dimensionality of the feature vector with a minimal influence on the accuracy. The best average accuracy of 87.1% was achieved using a feedforward neural network classifier with correlation measure to reduce the feature vector from 33 features down to 15.

Similarly, there have been other significant research works related to [83]. This includes the work by Bruździński *et al.* [85] which determined the breast cancer malignancy grade (high and intermediate) using a web-based automated classification system. They compared three segmentation methods, namely k-means, fuzzy c-means (FCM), and watershed. FCM was the most computationally intensive among the three studied segmentation methods and gave the most accurate nuclei segmentation results. The highest classification accuracy of 89.02% was recorded for the multilayer perceptron using a set of cellular and nuclear features extracted from LMI and HMI, respectively. The aim of including this study within the reviewed studies berceuse the best-achieved classification accuracy for malignancy grading problem was based on FCM segmentation results.

Additionally, based on the BR grading scheme and MLP, Jeleń *et al.* [86] developed a classification system for grading cancer malignancy. A high-performance result of 93.08% accuracy was achieved with an error rate of 13.5%. Further, Jeleń *et al.* [53] analyzed the role of nuclear segmentation from FNA biopsy slides and its influence on malignancy classification, by comparing three powerful segmentation approaches and testing their impact on the classification of breast cancer malignancy. Common classifiers like MLP, self organizing maps (SOM), PCA, and SVM were analyzed. Level Set Segmentation yielded the best results and led to a good feature extraction with the lowest average error rate of 6.51% over four different classifiers. The best performance was recorded for MLP with

an error rate of 3.07% using FCM segmentation. The usage of cell groups as a malignancy classification feature was analyzed by Jeleń *et al.* [87] by comparing discriminatory powers of calculated features and classifying error rates showing the feature discriminatory power on the classification. A cytological image segmentation process with FCM method was carried out by Krawczyk *et al.* [88], using an application of adaptive splitting and selection (AdaSS) ensemble classifier to design an efficient clinical decision support system for breast cancer malignancy grading. A dedicated ensemble model was used to exploit local areas of competence in the decision space to combat imbalanced classes in the dataset, resulting in better accuracy.

On the other hand, imbalanced datasets present significant challenges to the machine learning community. Typically, traditional classifiers may be biased towards the majority class which might lead to poor predictive accuracy over the minority class. Many attempts have been made to deal with unbalanced data in classification problems using different techniques such as data sampling, algorithmic level, and ensemble learning. Batista *et al.* [89] did a comprehensive investigation and evaluation of different existing methods that deal with the problem of class imbalance. The study provided evidence that the class imbalance problem does not systematically hinder the performance of learning systems; however, it is a critical problem for the classification task if overlapping characteristics are shared between classes. Krawczyk *et al.* [90] proposed a comprehensive, automatic clinical decision support system for breast cancer malignancy grading. The proposed system performs image analysis of biopsy slides using three different image segmentation methods (fuzzy c-means color segmentation, level set active contours technique and gray-level quantization method) with the aim of accurately segmenting cell nuclei regions for feature extraction purpose. Further, to handle the problem of an imbalanced dataset, an ensemble classifier named EUSBoost was used in this study. EUSBoost algorithm combines a boosting scheme with evolutionary undersampling to create a balanced training sets for each one of the base classifiers in the final ensemble. The aim of using the evolutionary approach was to select the most significant samples for the classifier learning step with respect to the overall accuracy performance. According to the study results, level-set active contours segmentation algorithm gave the highest discriminative power features. Further, the extracted features have shown that EUSBoost was able to outperform state-of-the-art ensemble classifiers when it used with an imbalanced dataset. Mazurowski *et al.* [91] examined the influence of data imbalance in simulated training datasets with the purpose of developing an ANN classifier for automated medical diagnosis systems. The authors constructed the ANN with two techniques involving classical backpropagation and particle swarm optimization (PSO). Based on the study results, it was concluded that even a low ratio of class imbalance in a training set could deteriorate the classifier performance. This is due to the fact that in real-environment data, there are other difficulties combined with the problem of imbalanced class distribution that can hinder the learning process of classification models such as

small sample size and data overlapping characteristics (samples from different classes share similar characteristics). Further, the study showed that backpropagation is better than PSO for imbalanced training data especially with small data samples and a large number of features. Guo *et al.* [92] conducted a study which included the combination of boosting and an ensemble-based learning algorithm. This algorithm generated the data to re-balance their original imbalanced dataset which included two classes. With their method, the samples with high correlation criteria (hard samples) from both the classes were identified during the boosting algorithm execution. A new independent synthetic dataset was generated for both classes and added to the original training dataset. Therefore, the distribution of the classes and the weights in the new training dataset were re-balanced using this methodology.

In recent years, deep learning-based methods such as CNN models have demonstrated impressive results for object detection and classification in a variety of domains including medical diagnosis [93, 94, 95, 96]. Recent studies in the digital pathology domain show the high accuracy of ConvNets-based frameworks for breast cancer diagnosis. Bejnordi *et al.* [97] applied automated machine learning techniques on 2387 digital images of benign and malignant breast biopsies to investigate mammographic abnormalities among 882 patients aged between 40 and 65 years. The proposed techniques accurately discriminated between stroma surrounding of invasive cancer images and stroma from benign biopsies with 0.962 area under the ROC curve. Rakhlin *et al.* [98] developed a computational approach based on deep convolution neural networks for breast cancer histology images. The authors have utilized several pre-trained deep neural network architectures including ResNet-50, InceptionV3, and VGG-16, as well as gradient boosted trees classifier to perform a four-class classification task (normal, benign, in situ carcinoma, and invasive carcinoma) and a two-class classification task to detect carcinomas or non-carcinoma cases. According to the study results, the accuracy of 87.2% was obtained from the four-class classification and 93.8% from two-class classification. Lao *et al.* [21] designed a case-based diagnosis approach for histopathological images using ConvNets. Their approach was able to make a diagnosis decision based on features learned in a combination of multiple magnification levels (40x, 100x, 200x and 400x). The study results showed that the case-based approach achieved better performance than the state-of-the-art methods when evaluated on the BreaKHis dataset. Recently, Żejmo *et al.* [99] presented a deep learning approach for automatic classification of breast tumors based on FNA biopsies to distinguish benign from malignant cases (malignancy diagnosis problem). The authors used two types of convolutional neural networks, i.e. GoogLeNet and AlexNet, in their proposed method. The method has been tested on cytological samples derived from 50 patients including 25 benign cases and 25 malignant cases and it started by dividing the cytological specimen of 200000×100000 pixels into small patches of size 256×256 pixels. Then, an SVM classifier was used to select the training and validation data

patches to guarantee the presence of a suitable amount of cell material in each selected patch. The best accuracy obtained was of 83% by the GoogLeNet model.

As discussed above, the first computer-aided malignancy grading system for breast cancer cytological images of FNA biopsies were presented by Jeleń *et al.* [83] based on adapting the histological Bloom-Richardson (BR) scheme [19] to grading cytological FNA slides. However, a cytological image of an FNA smear may be lacking the histopathologic features, such as a cellular structure, that a scheme like BR-scheme is based upon. The aim of the current research is to instead consider cytological grading systems such as Robinson's grading [36] as the basis of computer-aided CGSs for breast cancer FNAs. To achieve this objective, in this thesis, six computer-aided cytologic malignancy grading systems are proposed for FNA biopsies of breast cancer. Each one is designed to follow the cytological characteristics as defined in the six published cytological grading schemes (see Table 1).

Chapter 3

Computer-aided cytological grading systems for fine needle aspiration biopsies of breast cancer based on pathology-guided handcrafted features

There are two techniques which can be used for computer-aided diagnosis systems: (1) the traditional machine learning technique using handcrafted features, and (2) deep learning-based CNN. CNN has achieved high accuracy in different applications including medical image classification; however, monitoring meaningful features that reflect medical interpretations by CNN models is not possible. Furthermore, there is a linear relationship between the amounts of data required and the size of the training model. Thus, it is not recommended to use a CNN model when dealing with limited datasets because when a small dataset is fed to a model that has a large number of parameters such as CNN, it will fail to learn the patterns and easily overfit. On the other hand, the performance of the majority of the traditional classification algorithms requires an accurate image segmentation as well as handcrafted feature engineering. Handcrafted feature engineering results in highly correlated features related to the medical characteristics specific to the defined problem, which is not possible in CNN. Considering the above-mentioned issues with CNN, the upcoming sections of this chapter

focus on proposing computer-aided grading systems based on the traditional classification methods.

3.1 Automating the six well-known cytological grading systems of breast cancer

The first objective of this thesis was to propose six computer-aided cytological grading systems for FNA biopsy of breast cancer based on pathology-guided handcrafted features extraction. The six considered cytological grading schemas in this thesis are published in pathology-based studies. For these systems, to determine different sets of meaningful handcrafted features that reflect the biological behavior of cancerous cells, the features were selected to simulate, accurately and efficiently, the gradual malignancy change of the breast cancer cells as described by the cytological characteristics in the six considered cytological grading schemes (see Table 1). By doing so, the black-box aspect decision-making process of systems like those based on arbitrarily chosen features determined by convolutional neural networks is minimized [100]. The clinical adaptation of computer-aided diagnostic and malignancy grading tools is dependent on the transparency of the decision-making process which is hindered by the lack of interpretability of such automatic systems.

As mentioned in the literature review section, the first computer-aided malignancy grading system for cytological images of FNA breast cancer biopsies was presented by Jeleń *et al.* [83] based on adapting the histological BR scheme [19] to grading cytological FNA slides. However, a cytological image of an FNA smear may be lacking the histopathologic features such as the cellular structure that a scheme like BR is based upon. The aim of the current study is to instead consider cytological grading systems such as Robinson's grading [36] as the basis of computer-aided CGSs for breast cancer FNAs. To achieve this objective, in this section, six computer-aided cytologic malignancy grading systems are proposed for FNA biopsies of breast cancer, which are based on the six published cytological grading schemes (see Table 1). In this section, we give a full description of the proposed methodology for the grading systems and discuss each of their fundamental stages: Pre-processing, segmentation, features extraction, features selection, and classification. Further, we present the obtained experimental results and provide discussions and conclusions.

3.1.1 The methodology of the proposed frameworks

With respect to the first objective of this thesis, six CGSs for FNA biopsies of breast cancer were proposed based on handcrafted features. Different segmentation methods, features extraction, feature selection algorithms, and classifiers were considered and examined in the methodology to achieve our objective. Further, the sets of features that represent the cellular (from the LMI) and nuclear (from

the HMI) characteristics were calculated for the classification purpose. By converting the criteria of the CGSs into classification problems, the proposed frameworks were able to evaluate and assign a malignancy grade (G2 or G3) to an FNA slide. Entire details about the used dataset in this thesis were discussed in Appendix A. Five fundamental stages, detailed in what follows, are embedded in these CGSs, namely image pre-processing, segmentation, feature extraction, feature selection, and classification.

The flowchart of the proposed cytological grading systems is presented in Figure 18. The main objective of the proposed grading systems is to determine the malignancy grade of FNA biopsies of breast cancer. As an input, we feed a pair (each pair is processed separately) of images (LMI and HMI) belonging to a certain patient, to the CGSs. For the segmentation of the HMI, as shown in the flowchart, first, the images are enhanced by means of color deconvolution, contrast enhancement, and quantization processes in the pre-processing stage to yield better quality images. Second, we segment nuclei and cytoplasm regions from HMIs using a hybrid segmentation (GVF-MO) method and apply a nuclei filtration procedure to use only well-segmented nuclei for the feature extraction purpose in the segmentation stage. Third, different morphologic, pleomorphic and texture nuclear and cellular features are estimated. Next, an optimal subset of features is selected by Fisher's method for the classification purpose. For LMI, only the red channel images are extracted and a FCM algorithm is applied on this channel to segment these images. Finally, using the optimal subset of features from LMI and HMI, the systems are able to determine the malignancy grade (G2 or G3) of a given tumor biopsy (for either case or patient classification) in the classification stage. The lack of low malignancy (G1) images in this thesis is caused by the fact that these cases very rarely require FNA, and in recent years, there were only a few cases at the Medical University of Wrocław we collected the FNA biopsy images from.

Image pre-processing stage :

FNA slides are saved as three-channel RGB images, whose components are highly correlated and are mixed with different medical stains. Thus, it is difficult to obtain accurate image segmentation without performing some pre-processing techniques on those images. In this case, the FNA uses Hematoxylin-Eosin staining where Hematoxylin (blue) mainly stains the cell nuclei, and Eosin (magenta-red) stains cell cytoplasm. To aid in the image segmentation, as a pre-processing step, we perform color deconvolution as described by Ruifrok and Johnston [101] to determine the contribution or effect of the different stain levels in the images. Our goal was to extract the nuclei and cytoplasm layers by separating the contribution of the hematoxylin (nuclei) and eosin (cytoplasm) stains from the original images. Typically, the intensity values of the images have a non-linear

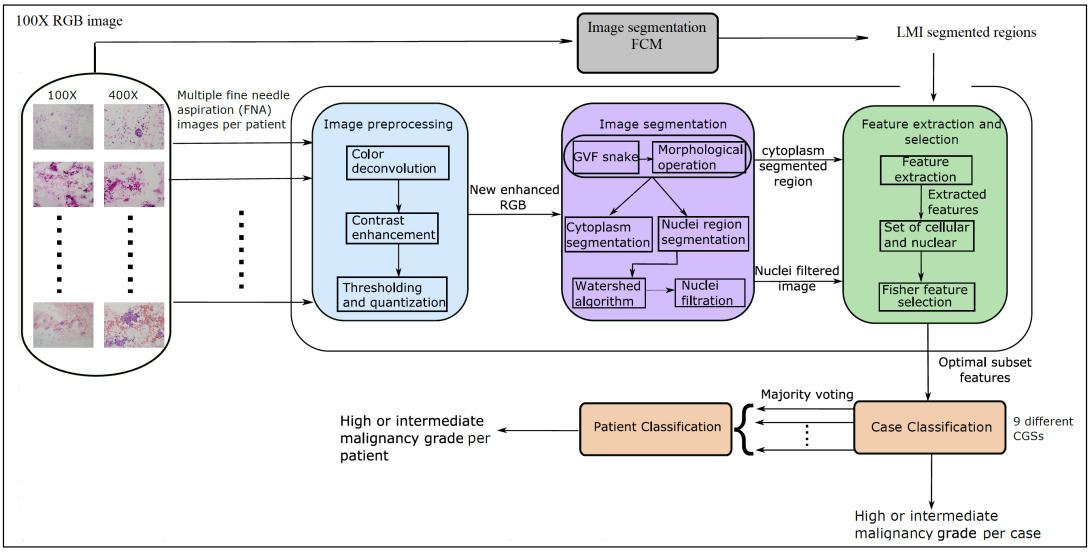


Figure 18: Overview of the workflow for the nine different cytological grading systems (CGSs).

relationship with the stain level. This implies that the stain level information cannot be directly separated from the RGB image. To solve this issue, the optical density (OD) matrix for each channel of RGB images was used because it has a linear relationship with the concentration of absorbing the material. The OD matrix is constructed from the amount of the stains and their absorption factors. Thus, it can be used to separate the contribution of multiple stains in a specimen, where each strain of the specimen is described by a particular optical density for the light in each of the three RGB channels. The OD matrix can be represented by a 3x1 OD vector defining the stain in the OD space of RGB channels. As exemplified in the matrix below (represents the OD matrix for the combination of hematoxylin, eosin, and DAB), the OD values of 0.18, 0.20 and 0.08 for the R, G, and B channels, respectively, represent the measurements of a sample stained with only hematoxylin stain.

$$\begin{array}{c} \\ \text{Hematoxylin} \\ \text{Eosin} \\ \text{DAB} \end{array} \begin{array}{c} \begin{array}{ccc} \text{R} & \text{G} & \text{B} \end{array} \\ \begin{pmatrix} 0.18 & 0.20 & 0.08 \\ 0.01 & 0.13 & 0.01 \\ 0.10 & 0.21 & 0.29 \end{pmatrix} \end{array}$$

In principle, to perform the color deconvolution, an orthogonal transformation of the above matrix has been taken, but before that, the OD matrix must be normalized in order to balance the absorption factor of each stain accurately by dividing each OD value by the length of the stain. As an example of the above normalized OD matrix can be seen below:

$$\begin{pmatrix} 0.65 & 0.70 & 0.29 \\ 0.07 & 0.99 & 0.11 \\ 0.27 & 0.57 & 0.78 \end{pmatrix}$$

Next, the inverse of the normalized OD matrix was used to separate the contribution of each stain color in the image. The process of separating the colors is known as color deconvolution. To get the finalized stained layers, the inverse of the OD matrix is multiplied with the image (see Figure 19). The following matrix represents the color deconvolution matrix corresponding to the above example of the color matrix:

$$\begin{pmatrix} 1.88 & -0.07 & -0.60 \\ -1.02 & 1.13 & -0.48 \\ -0.55 & -0.13 & 1.57 \end{pmatrix}$$

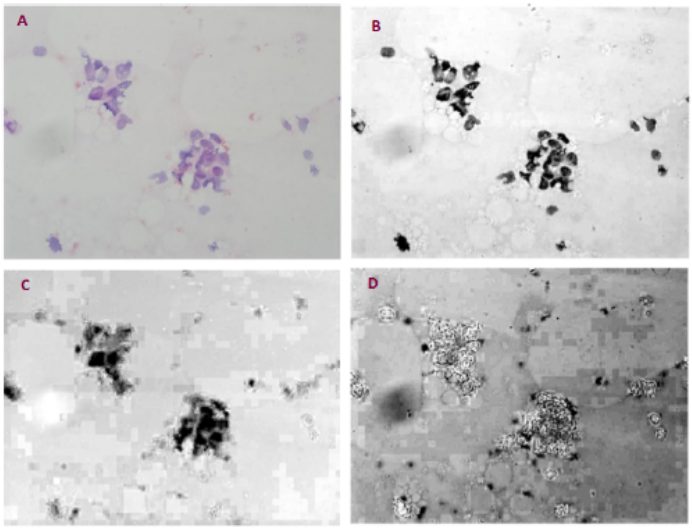


Figure 19: Example images of the color-deconvolution process for an intermediate malignancy case from JELEN16 dataset. (A) original image, (B) Hematoxylin layer, (C) Eosin layer and (D) DAB layer.

To improve the quality of the extracted Hematoxylin layer (nuclei layer), we adjusted the intensity values using contrast enhancement. A multi-level threshold was computed for the adjusted image using Otsu's method; then, the quantization process was applied using the estimated threshold values to segment the image into three regions represented by distinct labels. The resulting labeled image was converted into a color RGB image for the purpose of visualizing the labeled regions. A representative example of the results obtained from these three steps is shown in Figure 20. Finally, for segmentation, the active contour method (GVF-snake), where the segmentation method

consists of GVF-snake and MO, and which requires an initial curve was used. In this context, the boundaries between the regions in the green and blue channels of the extracted image were used as initial contours to extract nuclei and cytoplasm regions, respectively (see Figure 21).

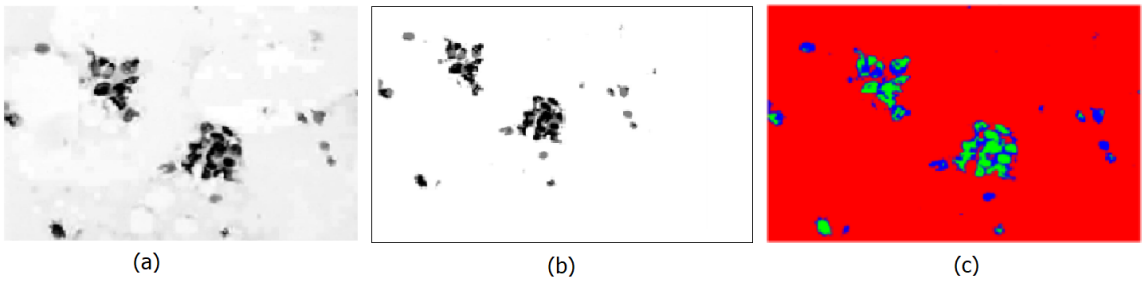


Figure 20: Example images of the pre-processing steps for the hematoxylin channel image: (A) contrast-enhanced image, (B) labeled image by quantization process and (C) colored RGB image.

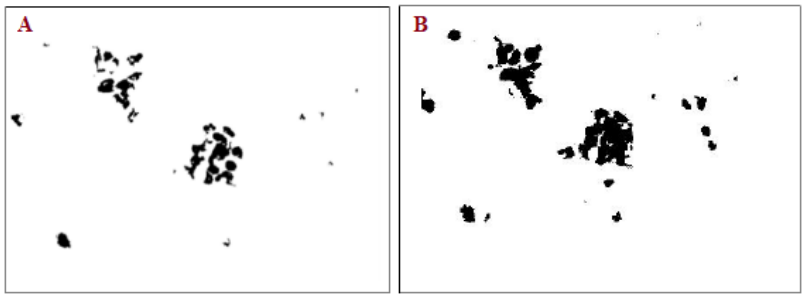


Figure 21: Example of the extracted green and blue channel images that were used as initial contours with the GVF-MO: (A) represents the green channel image and (B) represents the blue channel image.

Image segmentation stage :

Precise image segmentation is a key factor in achieving an accurate classification result. Medical image segmentation is one of the most challenging and essential tasks in most computer-aided diagnosis systems. This is due to large complexity and variability of appearances and shapes of image objects (inhomogeneous in nature) [102]. Cytological images originating from different sources may differ significantly due to their imaging method or smear preparation. The low quality of images caused by the use of a different camera, microscope lighting conditions, and staining material makes the segmentation difficult (see Figure 22). In addition, different and overlapped objects (nuclei, cytoplasm, red blood cells) inside the image (see Figure 23) further complicate the process [103]. In order to automate the image segmentation process and to achieve more accurate segmentation results, different sophisticated segmentation approaches have been examined in this study to perform nuclei isolation.

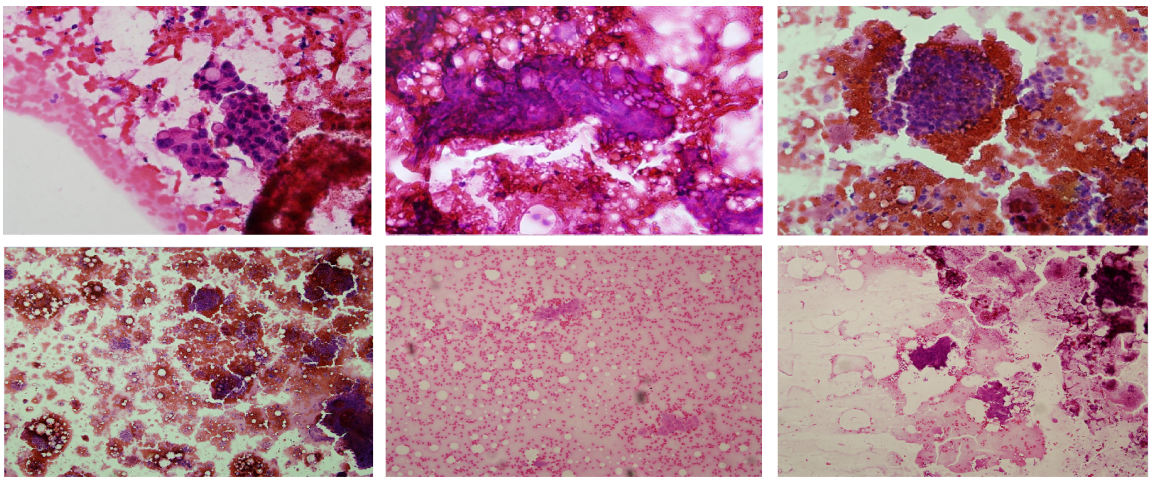


Figure 22: Examples of low quality HMIs of `JELEN_MERGE01` dataset.

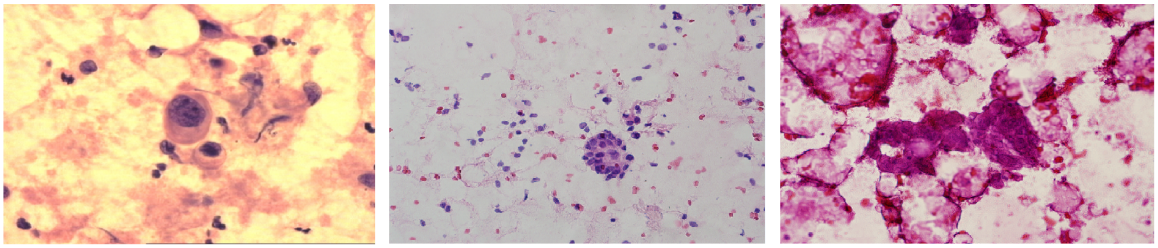


Figure 23: Examples of different and overlapped objects in the cytological images from the `JELEN_MERGE01` dataset, (a) different sizes and shapes of nuclei, (b) different nuclei, RBC and cytoplasm regions, (c) overlapped nuclei regions.

As mentioned earlier, two magnification images (low with 100x size and high with 400x size) have been used in this study. Thus, for the LMIs, a FCM clustering algorithm was used for the segmentation of clusters and single nuclei regions (see Figure 24 and Figure 25). For the HMIs, according to our experimental results, a hybrid segmentation method, namely a combination of the GVF-Snake algorithm and morphological operations (MO) (process an image based on shapes) which we designated as GVF-MO, was selected to segment these images, giving more precise nuclei and cytoplasm boundary results compared to other examined methods such as the level set (LS) method [104] and FCM. Though the Snake algorithm was originally proposed by Kass *et al.* [105] a modified version was introduced by Xu *et al.* [106] due to the limitation in the traditional snake which it required the initial contour to be placed close to the object to prevent it from converging to a local minimum. Thus, the GVF-snake handles this problem because it includes a new energy field that has the desired properties of a large capture range as well as the existence of an energy that guides the initial contour to detect boundary concavities [107]. The GVF-snake was initialized outside the boundaries of the nuclei; hence, it didn't pass the strong edge of nuclei boundaries to penetrate the boundaries of nuclei and this algorithm combined with morphological operations reached the same contours of the nuclei for most of the segmented images. The aim of adding the MO is to fill holes (a set of pixels) in the segmented nuclei by adjusting the nucleus pixels based on the neighborhood pixel values.

On the other hand, the LS approach shows good segmentation results for nuclei boundaries, but it failed to accurately segment some images due to the low quality of these images. Thus, this algorithm was unable to pass the boundaries in order to reach inside most of the nuclei regions. Further, this algorithm produced double boundaries instead of the desired single boundary for each nucleus as shown in Figure 26. With regards to the FCM results, due to variations in the nuclei sizes and shapes in the cytological images, tuning the algorithm was challenging and it was difficult to obtain accurate results. Moreover, it is very time-consuming if the data is very large as the case with a large number of nuclei existing in most of the cytological images. Thus, estimating the membership matrix was time-consuming as the boundaries of each nucleus had to be segmented individually. Therefore, most of the obtained nuclei regions were clusters of connected nuclei with irregular boundaries (see Figure 27). As a result, the GVF-MO method was selected for nuclei segmentation because it showed the best results for most segmented images. Two phases of segmentation are involved at this stage: Nuclei and cytoplasm segmentation.

Nuclei segmentation: In the first stage of nuclei segmentation, we use GVF-MO (defined above) to segment the nuclei regions. A parametric active contour or snake is a powerful segmentation method used to achieve sub-regions with continuous boundaries. It is a curve move based on internal

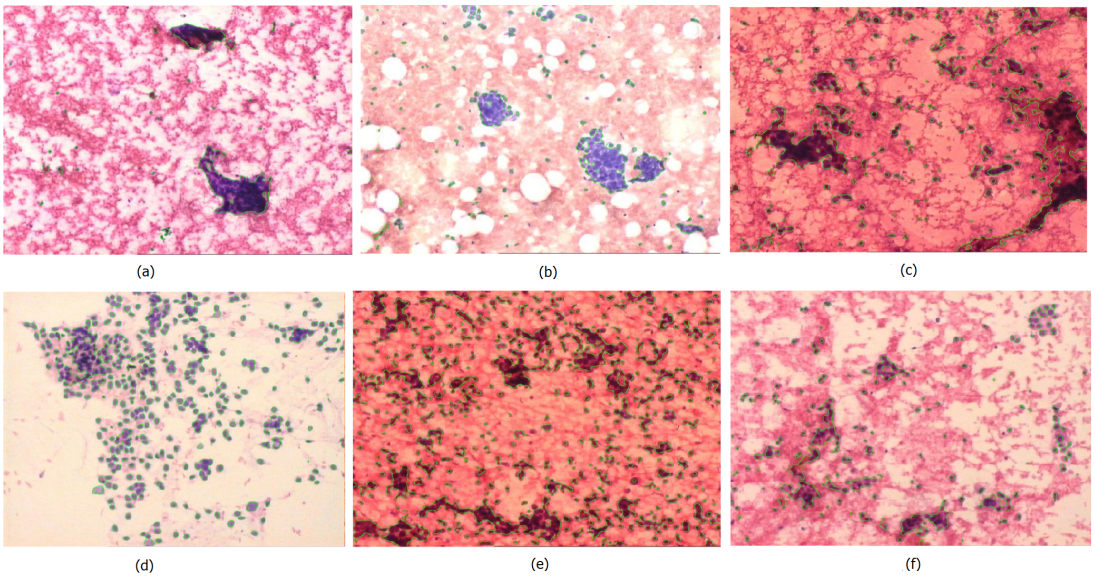


Figure 24: Examples of the final segmented boundaries for the LMI obtained from FCM approach, (a-c) belongs to G2 samples, (d-f) belongs to G3 samples.

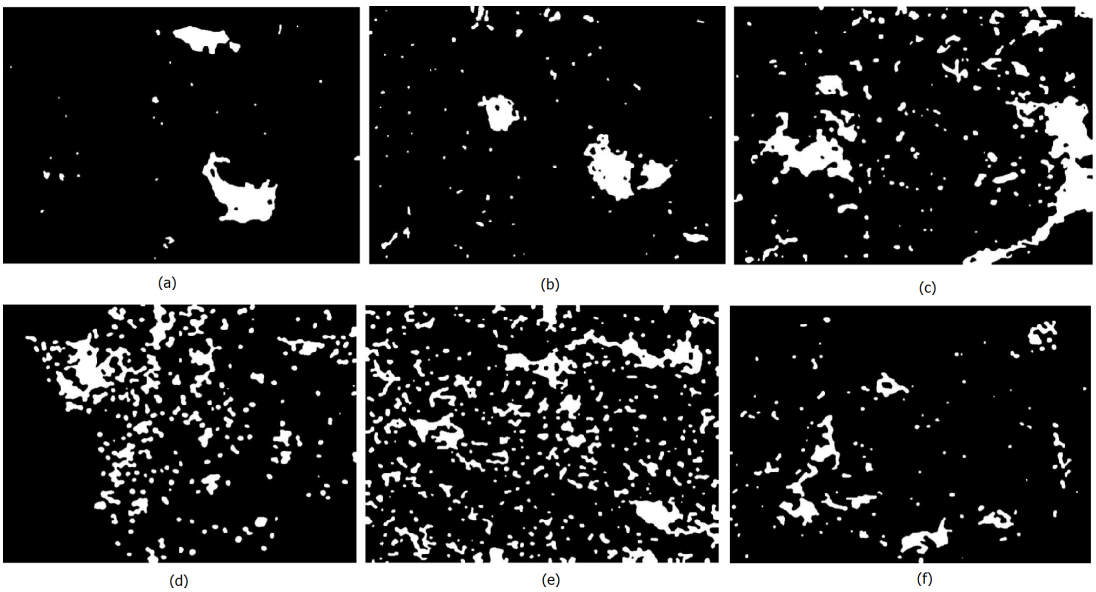


Figure 25: Examples of the binary images for the LMI obtained from FCM approach, (a-c) belongs to G2 samples, (d-f) belongs to G3 samples.

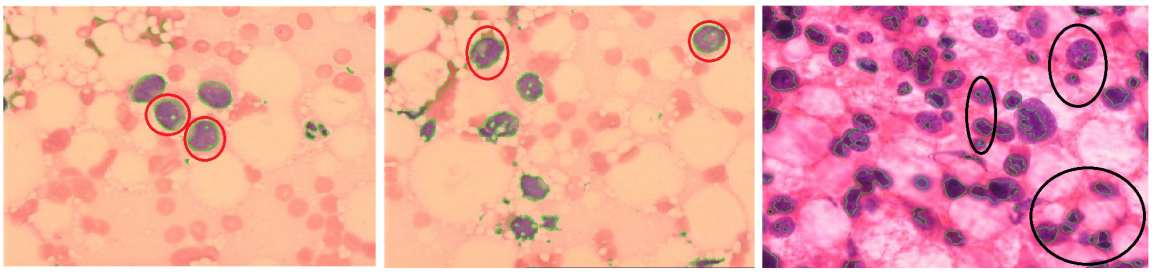


Figure 26: Examples of the final segmented boundaries for nuclei regions of HMI belongs to G2 sample obtained by LS approach. The red borders in (a) and (b) images highlight examples of double nuclei boundary results. The black borders in (c) image highlight examples of inaccurately segmented nuclei boundaries (did not fully capture the actual nuclei boundaries. The double boundaries are highlighted in green).

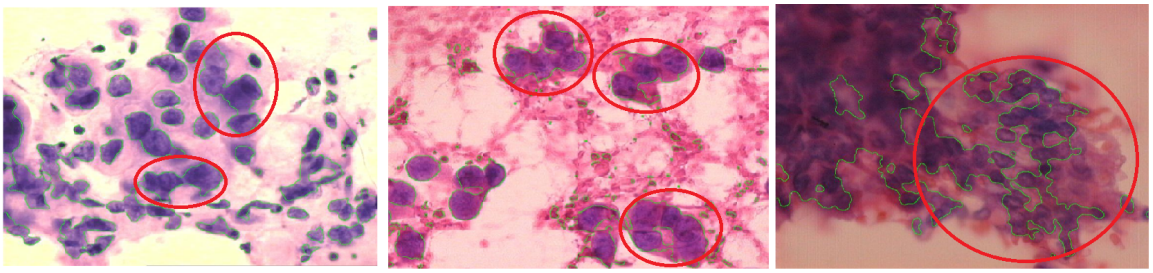


Figure 27: Examples of the final segmented boundaries for nuclei regions of HMI belongs to G3 sample obtained by FCM approach. The red borders in the images highlighted the results of the example of inaccurately segmented nuclei boundaries. The FCM results illustrate a continuous boundary (highlighted in green) that encapsulates several cells or groups of cells rather than boundaries of individual cells.

snake energy that forces the curve to be smooth and continuous, and external image energy that forces the curve to search for desirable image properties such as lines, edges or object boundaries. The main concept of the active contour or snake algorithm is the use of initial boundaries (extracted green or blue channel images in this thesis) represented by closed contours. Then, it iteratively modifies the contours by applying size shrinking (it meets the edge of the objects) or expansion processes to the contours according to the mentioned internal and external energies (estimated from the image). The shrinking or expansion processes represent the evolution of the contours and perform based on energy function minimization. The snake's convergence occurs when the internal and external energies interact where the minimum energy is located. The method intends to minimize the following energy function:

$E_{snake} = \int_0^1 [E_{int}(v(s)) + E_{image}(v(s))]ds$

Where E_{int} represents the internal energy of the snake and E_{image} represents the external energy of the image.

As we mentioned above, due to traditional snakes' limited ability of handling only simple structure objects and their sensitivity to their initial conditions (contour must be placed close to the object), the GVF-snake was used to address these limitations. In GVF-snakes, the gradient vector flow field $F_e xt = V(x, y)$ is derived from the following energy function:

$E_{snake} = \int\int (\mu(u_x^2 + u_y^2 + v_x^2 + u_y^2) + \nabla f^2 |V - \nabla f|^2)dxdy$

Where V(x,y) is the field vector flow of image, f is the edge map and μ is a decreasing function of the gradient magnitude.

In addition to the use of GVF-snake, MO have been used in this stage to correct the boundaries and fill the holes in the segmented nuclei regions. This method (GVF-MO) provides a powerful interactive tool for image segmentation (see Figure 28).

As a second stage of the nuclei segmentation, we applied a watershed algorithm to separate individual nuclei from clusters of connected nuclei that were indistinguishable to the GVF-MO. Although good segmentation results from the GVF-MO method, it cannot separate touching nuclei where the outputs obtained from the first stage of nuclei segmentation were mixtures of separate nuclei regions as well as clusters of connected nuclei that were assumed to be large nuclei by the GVF-Snake. Figure 29 shows an example of the initial nuclei segmentation results where there are clusters of connected nuclei as well as individual nuclei regions outputted by the GVF-MO segmentation method. To re-segment the clusters of connected nuclei and increase the number of separated nuclei for the feature extraction purpose, we utilized a watershed segmentation algorithm (powerful tool to handle such

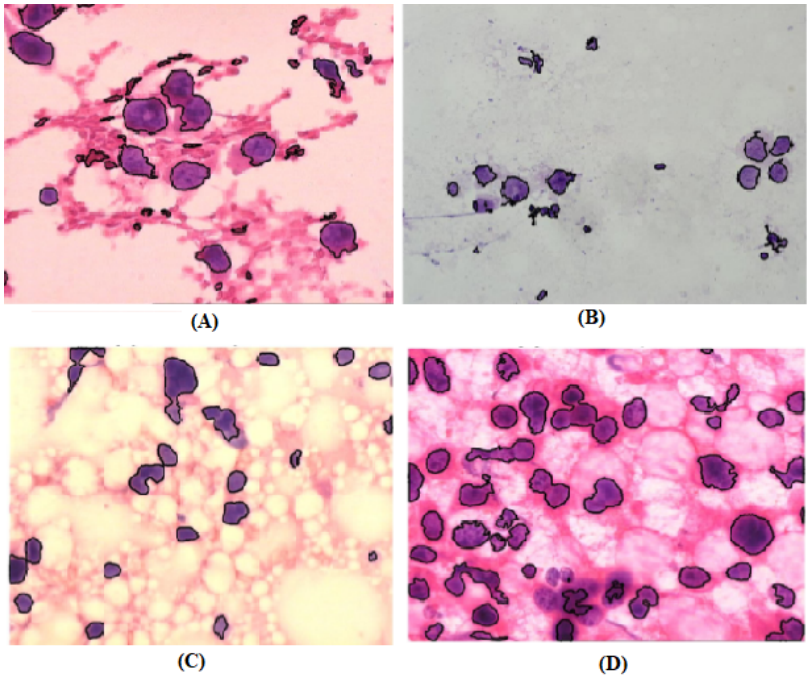


Figure 28: Examples of final segmentation results by GVF-MO for HMIs of the `JELEN_MERGE01` dataset: (A) represent an intermediate malignancy, while (B), (C) and (D) correspond to high malignancy.

a problem) based on mathematical morphology watersheds as proposed by Malpica *et al.* [108] to separate touching nuclei (clusters) in the images. This step of separating clusters of connected nuclei led to the improvement of the overall segmentation results because it produced enough separated nuclei for feature extraction purposes. This is due to the fact that obtained results from the GFV-snake for some of the images were only clusters of connected nuclei, and for some other images, a few individual nuclei and clusters of connected nuclei which required the second stage of nuclei segmentation (see Figure 30).

The watershed algorithm effectively segmentd the clusters of connected nuclei regions into individual nuclei based on the nuclear size (NS). Specifically, we examined the nuclei size criterion and supposed that if the nuclei sizes are bigger than a specific, experimentally determined NS, we assumed that those sizes belong to clusters of nuclei. So, in stage two, we applied the watershed segmentation algorithm to re-segment those nuclei clusters to individual nuclei (see Figure 31).

The algorithm works based on finding catchment basins (nuclei regions) and watershed lines in an image by considering the image as a surface in which light pixels are high and dark pixels are low. Thus, the algorithm focuses on marking nuclei regions and background locations. To

achieve this goal, firstly, we computed the distance transform (binary image), that represents the distance from every pixel to the nearest nonzero-valued pixel, of the image's complement. Secondly, we negated the computed distance transform to convert the bright areas into catchment basins. Thirdly, we computed the watershed transform for the resulted image that finds the catchment basins or watershed ridge-lines in the image. The watershed transform is the so-called label matrix in which positive integer values are the locations of each nucleus. Finally, we used the zero-valued elements of the estimated label matrix, that are located along the watershed lines, as a mask to separate the nuclei in the original image.

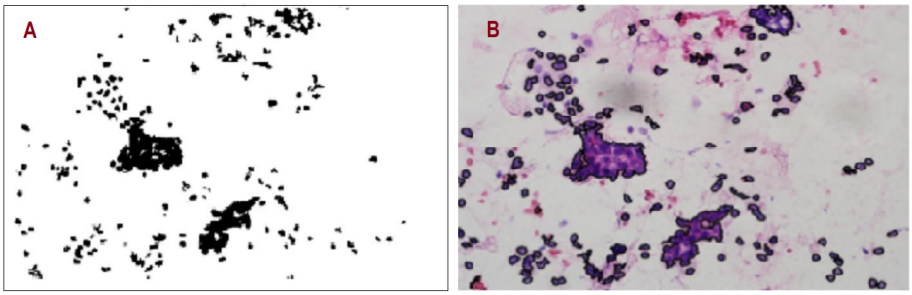


Figure 29: Initial segmentation results using GVF-MO method for intermediate malignancy case of JELEN16 dataset: (A) binary image produced by applying the segmentation process and (B) segmented boundaries of nuclei on the original image.

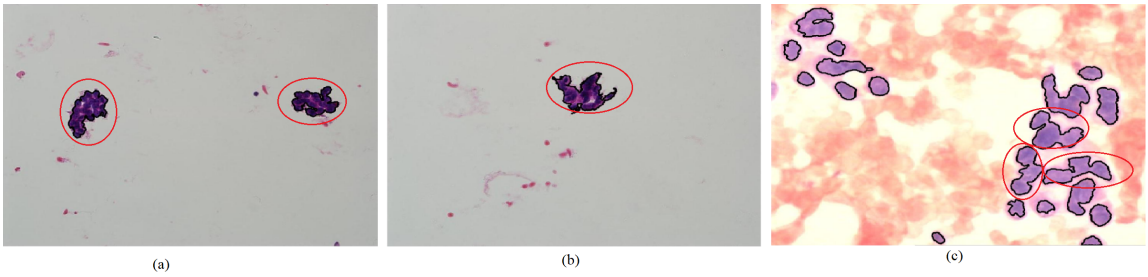


Figure 30: Examples of segmented images by GVF-MO which required further nuclei segmentation step to separate the clusters into individual nuclei in (a) and (b) images where the highlighted parts represent examples for only clusters of connected nuclei existing in the image, while in (c) image, the highlighted parts represent examples of few numbers of single nuclei existing in the image.

To summarize the nuclei segmentation task using the green channel image obtained by the color deconvolution process, we first cleaned the border components by removing all connected components that touch any border of the image to faster the segmentation process. The next, we adjusted the contrast of the green channel image. Then, we applied multilevel thresholds using Otsu's method and applied a quantization process on the basis of multilevel thresholds to segment the image into three regions by distinct labels. Finally, we converted the three labeled regions image into three colors

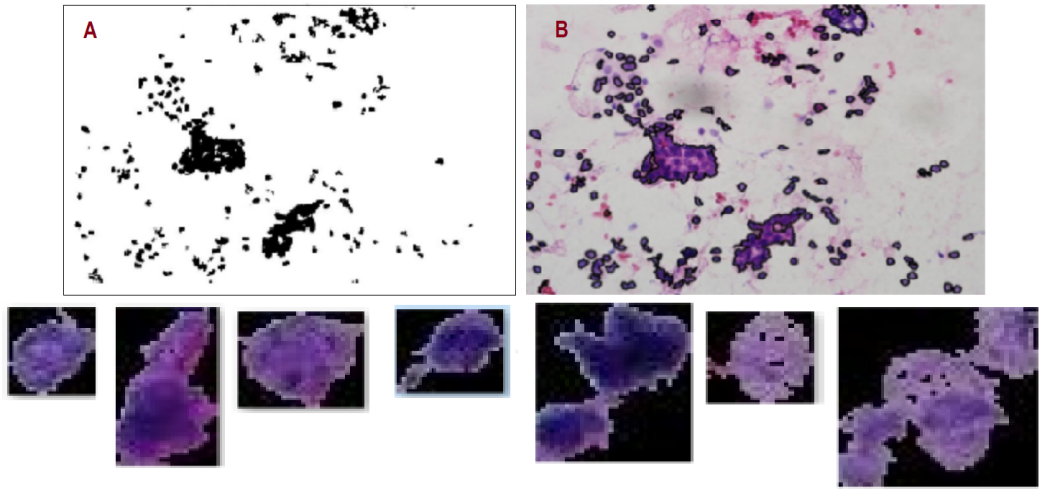


Figure 31: Examples of individual nuclei obtained from re-segment the above clusters by the watershed algorithm.

RGB image to achieve a better visualization. All these steps were performed as a pre-processing stage to aid the image segmentation task. To perform nuclei segmentation, we fed the enhanced image to GVF-MO method where the GVF-snake algorithm accurately isolated most of the nuclei regions as the first step of the nuclei segmentation task. Then, we removed small cell nuclei that have size 15 pixels or less (to remove noisy nuclei regions) that helped in tuning the segmentation method to isolate most the reasonable nuclei sizes in the images for better feature extraction results. Then, we filling holes of rough nuclei and corrected the nuclei boundaries using MO to get smoother nuclei regions. In the second stage of nuclei segmentation, we applied the watershed algorithm, as some clusters of connected nuclei didn't separate into individual nuclei regions when they were assumed to be large size nuclei by the GVF-Snake. The watershed algorithm separated these clusters into individual nuclei based on NS criterion (NS=100 was selected experimentally as threshold value to resegment the clusters of nuclei). Thus, the watershed algorithm maintained the nuclei size with 100 pixels or less; if the nuclei size was greater than 100 pixels, the algorithm would assume the nuclei belonged to a cluster of connected nuclei, and would therefore resegment the cluster into individual nuclei.

Nuclei filtration: In the cytological images, some of the nuclei clusters were not completely re-segmented following the application of the watershed algorithm due to the poor quality of some of the images. Further, the presence of red blood cells, and false-positive results caused by geometric arrangements in the background were incorrectly identified as nuclear boundaries. To avoid this problem, at this point, we mimicked the process of a pathologist seeking out nuclei regions with

clear structures and used their expert knowledge to evaluate their characteristics and filter the determined nuclear regions. The results from the watershed segmented nuclei were classified as well-segmented (used for feature extraction and classification purpose) or poorly-segmented (ignored results) nuclei. To perform this task, we adapted the nuclei filtration procedure as proposed by Filipczuk *et al.* [52], who applied their procedure to the results of the CHT, where three features were estimated from segmented regions and used to classify these regions as correct or incorrect nuclei regions, whereas we adapted the procedure to the results of the watershed algorithm using specific nuclear morphological features. To do so, first, for all the segmented nuclei regions, seven nuclear features were calculated: Euler number (the total number of pixels in a nucleus minus the total number of holes in that nucleus), size (estimated based on the length of a nucleus), entropy (measures the nucleus disorder), homogeneity (the closeness of the distribution of nucleus pixels), energy (measures the nucleus uniformity), correlation (how correlated a pixel is to its neighbor of a nucleus) and standard deviation (uniformity of nuclei). An SVM classifier, with a Gaussian radial basis kernel function in which the scale of the Gaussian kernel was decided heuristically, was trained using these features to classify the segmented regions as well or poorly-segmented (including joined or overlapped nuclei and false positives) nuclei. The used dataset was prepared manually and contained 2211 nuclei regions, 1273 of which were well-segmented nuclei, and the remaining 938 classified as poorly-segmented. Using the JELEN_MERGE01 dataset (defined in Appendix A), the obtained classification accuracy was 80.24% evaluated using the 5-fold cross-validation technique. Figures 32 and 33 show representative examples of the well and poorly-segmented nuclei region results. Though we obtained a misclassification rate of 20%, the classified individual well-segmented nuclei regions from each image were sufficient to classify the image as belonging to an intermediate (G2) or high (G3) malignancy grade.

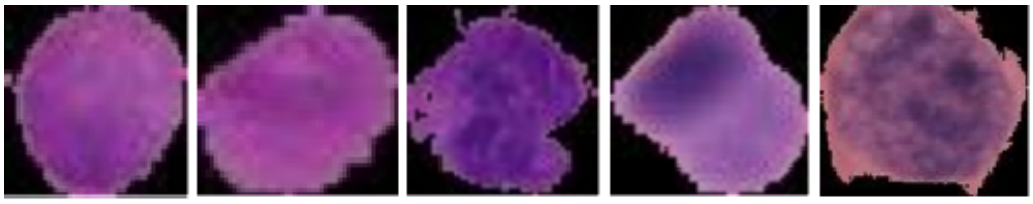


Figure 32: Example of well-segmented nuclei for intermediate malignancy case from the JELEN08 dataset.

Cytoplasm segmentation: In this phase of the segmentation, the combination of GVF-MO was used to segment the cytoplasm regions using the hematoxylin layer images on the base of the blue channel images. The cytoplasm segmentation was applied in order to mimic cytoplasm characteristics, as in some of the cytological schemes (Fisher's [34] and Taniguchi *et al.* [39] grading

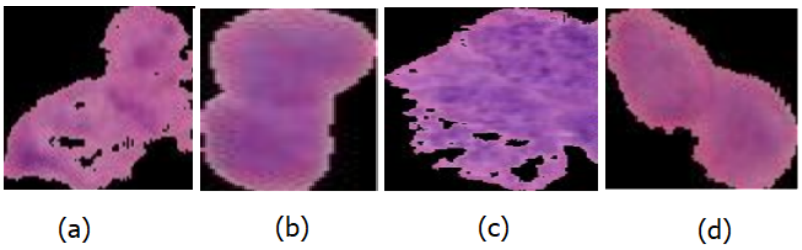


Figure 33: Example of poorly-segmented nuclei results for intermediate malignancy case from the JELEN08 dataset: (a and c) false positive results and (b and d) overlapped nuclei.

schemes). In other words, cytological criteria such as nuclear chromasia (measures the nuclear density or color) and NCR (the ratio of the nuclei size to the cytoplasm size of a cell) are required to examine the characteristics of the cytoplasm. The cytoplasm segmentation begins with the color separation, quality enhancement and quantization procedures for images as explained in the nuclei segmentation stage. Finally, the blue channel images of the new colored RGB images (originally of hematoxylin layer images) are fed into GVF-MO as initial contours for the cytoplasm segmentation purpose. The well-segmented nuclei, as well as the segmented cytoplasm regions, are used in the next stage of feature extraction to estimate sets of different cellular and nuclear features. Figures 34 and 35 show examples of the final nuclei and cytoplasm segmentations of G2 and G3 samples, respectively.

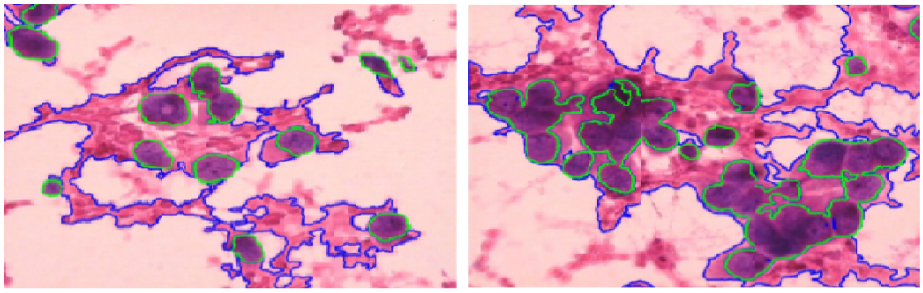


Figure 34: Final cytoplasm and nuclei segmentation results using GVF-MO for intermediate malignancy case from the JELEN08 dataset, where the green color boundaries represent nuclei regions, while the blue color boundaries correspond to cytoplasm regions.

In summary, the GVF-MO segmentation method was selected experimentally and applied for two tasks, namely nuclei segmentation and cytoplasm segmentation.

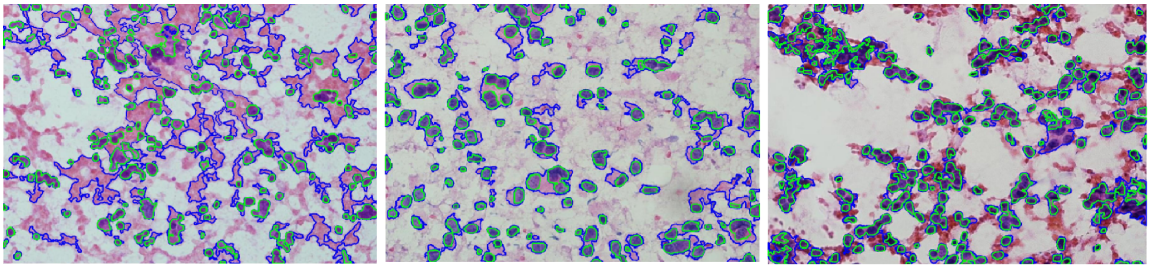


Figure 35: Final nuclei and cytoplasm segmentation results using GVF-MO for high malignancy case from the JELEN16 dataset, where the green color boundaries represent nuclei regions, while the blue color boundaries correspond to cytoplasm regions.

Handcrafted feature extraction stage:

For most computer-aided diagnosis systems, feature extraction is as important as image segmentation. The choice of more relevant handcrafted features yield good classification results. This section elaborates on the extraction of different sets of cellular and nuclear features that reflect the biological behavior of cancerous cells using the segmented nuclei and cytoplasm regions from HMI as well as the segmented regions from the LMI. The feature extraction involves two phases. In the first phase, to reflect the cell dissociation and cellular characteristics criteria that are respectively present in Robinson's and Mouriquand's schemes, a set of three structural features (area of groups, number of groups and dispersion) were extracted from LMIs (100x) for each case. These features were able to characterize the cells' ability to form clusters or to disperse within the image (see Figure 36) as proposed by Jeleń *et al.* [83].

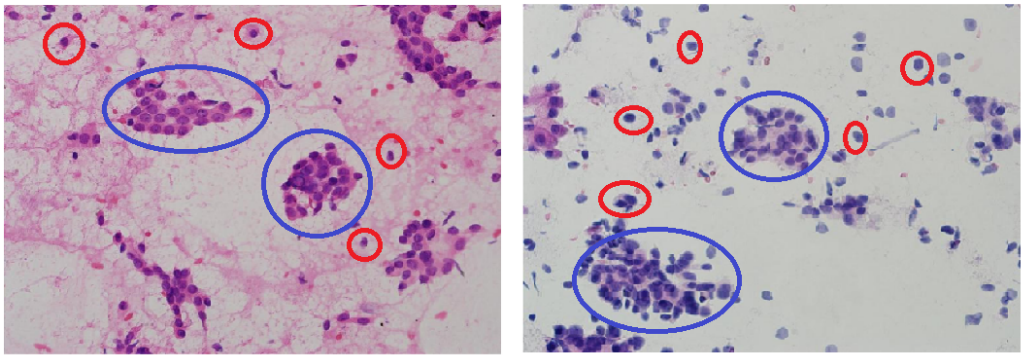


Figure 36: Example images belong to G2 samples from the JELEN16 dataset illustrate cluster and individual nuclei regions. The blue outlines highlight the clusters regions while red borders highlighted the individual nuclei regions.

In the second phase of the feature extraction, different nuclei characteristic features, that are required in all six developed grading frameworks, were evaluated. To achieve this goal, this phase was divided

into three stages to estimate three different sets of nuclear features. These features are derived from HMIs (400x) only and represent the nuclear pleomorphic, textural, and morphologic characteristics. The aforementioned features are capable of providing accurate information about the shape, size, and staining information of cell nuclei. The five nuclear pleomorphic features [83, 103] are estimated to reflect the nuclear size, anisonucleosis (morphological manifestation of nuclear injury), cellular pleomorphism, nuclear pleomorphism, and nuclear feature malignancy factors (see Figure 37) of the discussed grading schemes. These features reflect the variance in the shape and size of cells and their nuclei as shown in Figure 37.

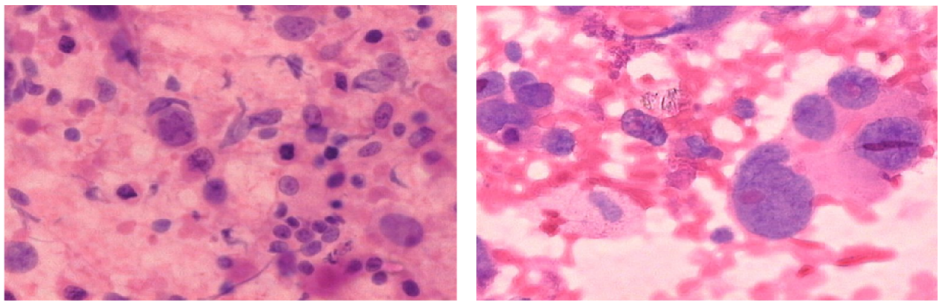


Figure 37: Examples of variation in size, shape and margin of cell nuclei of G2 samples from the JELEN08 dataset.

On the other hand, in order to calculate nuclear textural features, gray-level co-occurrence matrix (GLCM) and gray-level run-length matrix (GLRLM), as used by Filipczuk *et al.* [52], were estimated. The GLCM matrix describes how often different combinations of pixel grey levels occur in an image. The GLRLM matrix gives the size of homogeneous runs for each gray level. The extracted textural features (the full list of these features was displayed at the end of this section) based on these two matrices were able to reflect the characteristics of nucleoli, nuclear chromatin, nuclear chromasia, chromatin pattern, and chromatin granularity malignancy factors (interpret the hyperchromatism or very dark nuclei) of these grading schemes (see Figures 38 and 39). A total of 21 features were estimated: 10 textural features from the first matrix and 11 from the second.

The last estimated group of features was a set of 16 nuclear morphologic features that reflected the nuclear margin (measures the irregularity of the nuclear membrane), cell uniformity (measures the similarity in size and shape of cell structure), nuclear membrane (measures the irregularity of the membrane that encloses the nucleus and is another indicator of the nuclear margin), and cellular size. These features describe the irregularity or variance of the size, shape, and margin of cells and their nuclei. The mean and variance were then calculated for each of the mentioned features (pleomorphic, textural and morphologic) resulting in a total of 84 different nuclear features. These 84 nuclear polymorphic, textural and morphologic features have been used with all cytological grading

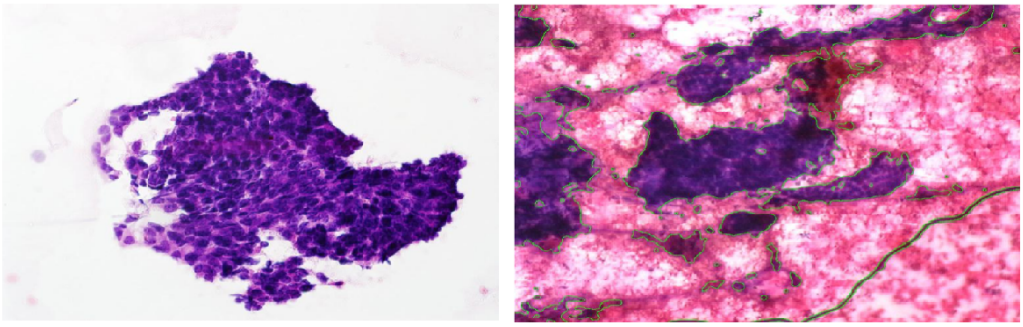


Figure 38: Examples of hyperchromatism or highly pigmented (dark purple to black) nuclei regions belonging G2 samples from the `JELEN_MERGE01` dataset. The dark staining or hyperchromatism (very dark purple-black regions illustrated above) indicate an increase in DNA; in other words, visible abnormalities in the nuclei.

schemes (Table 1), except for Howell's scheme [37] which only included the nuclear polymorphic features out of the three estimated features sets, one mitosis feature, and three low magnification cellular features. The other estimated features of this scheme were the MCC and the three cellular features estimated from the LMIs.

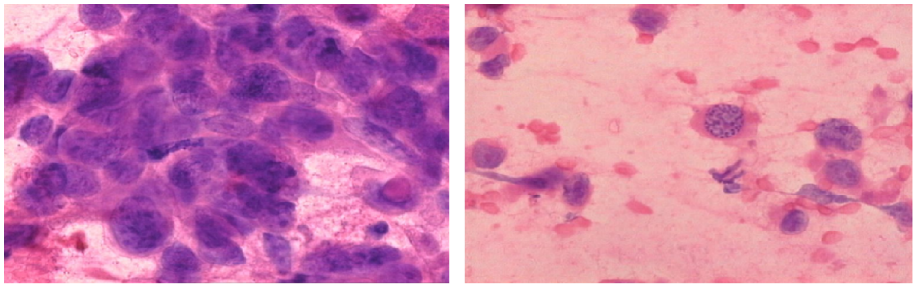


Figure 39: Examples illustrating the highly pigmented (granularity) of chromatin (being granular) in cell nuclei belong to G2 samples from the JELEN08 dataset. Highly pigmented, clumped granules in chromatin within the nucleus indicate cell mutation and malignancy criterion.

Due to limitations associated with the cytological images, the malignancy criteria that involved these cytological schemes (Table 1) were modified to facilitate the implementation of these schemes. One of the biggest challenges during the feature extraction stage was related to the estimation of certain cytological criteria, including the CN, cellular characteristics, MC and NTCN. To begin with, CN is a form of cell injury due to external factors (infection, toxins, or trauma) which lead to type-T3 cell death due to autolysis [109]. From a medical point of view, it is defined as an irreversible loss of plasma membrane integrity [110]. Mostly, this loss of the integrity of cell membranes exists in the cytological images due to the disintegration of tissue and cell structures during the material

extraction and FNA slides staining (See Figure 40). Therefore, we sought to find efficient features that could be used to reflect the mentioned criteria using the cytological images. To achieve this goal, several relevant research papers were reviewed. In the study of Alvarez *et al.* [110], the CN has been defined based on three significant morphological criteria, namely plasma membrane rupture, dilatation of cytoplasmic organelles and loss of nuclear and cytoplasmatic details of a cell.

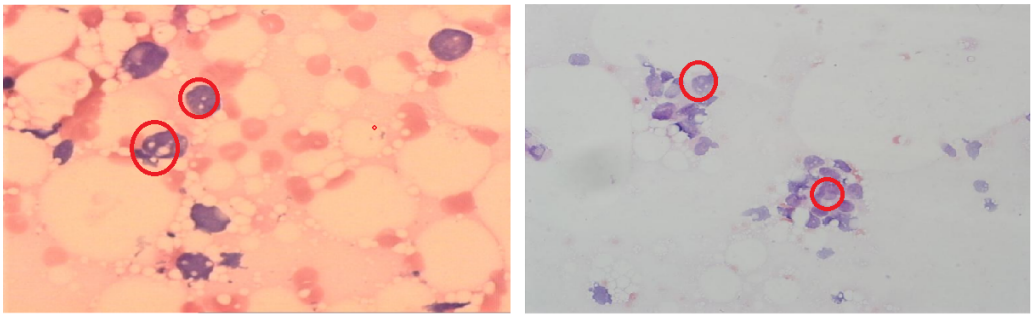


Figure 40: Examples of necrosis cell images in select G2 samples taken from the `JELEN_MERGE01` dataset. The holes within the cells, outlined in red, indicate the possibility of cell necrosis. .

In contrast, Rello *et al.* [111] proposed different morphological criteria (apoptotic cell rounding and shrinkage and the appearance of membrane bubbles in early CNC necrosis) to distinguish cell death mechanisms. Meanwhile, Kroemer *et al.* [112] proposed a comprehensive study to discuss the classification of cell death based on four different criteria: morphological, enzymological, functional and immunological. The study concluded that although the morphological criteria are mostly used to estimate different types of cell death, there is still a need to find the best biochemical mechanism that would classify the cell death types accurately.

In addition, CNC itself consists of several stages that start with a cell shrinking process followed by the extension of the plasma membrane and the separation of cell fragments into apoptotic bodies. Thus, a comprehensive study is required in order to discuss the automated detection of cell death by CNC in breast cancer. Further, samples of the FNA biopsy images should be previously prepared to examine the CNC and its types as they provide a good indicator of the cytological tumor grading problem. Given these difficulties and limitations in detecting the CNC malignancy factor in cytological images, CNC was ignored in this thesis.

Another major challenge was related to some of the cellular characteristics of malignant tumors such as lack of differentiation, known as anaplasia (see Figure 41); pleomorphism, which reflects the variety of change in size and shape of cancerous cells (see Figure 37) and their nuclei; and, irregular chromatin within nuclei (see Figure 42). Typically, to estimate these criteria, efficient segmentation of cell boundaries is required; a similar approach was employed for nuclear segmentation (nuclei

segmentation means the process of segmenting the center part of a cell which differs from cell segmentation, which segments the whole cell structure). However, this task seemed to be very challenging to accomplish with the cytological images due to the presence of similar pixel intensity values belonging to the cell membrane and background regions. Moreover, most of the cell structures are get destroyed during the material extraction and staining procedure. Hence, the cellular criteria were estimated in terms of their nuclei characteristic variability. In other words, the cellular criteria were evaluated based on the corresponding nuclei change that occurs due to the transformation of normal cells into cancerous cells in malignant tumors. This assumption is based on the fact that the biological behavior of any cancerous cell is usually observed first in its nucleus. In the reviewed pathology-based studies [32, 113], the authors have estimated the nuclear size instead of the cell size using Robinson's grading schema, due to the limitation of cell structures in the cytological slides of FNA.

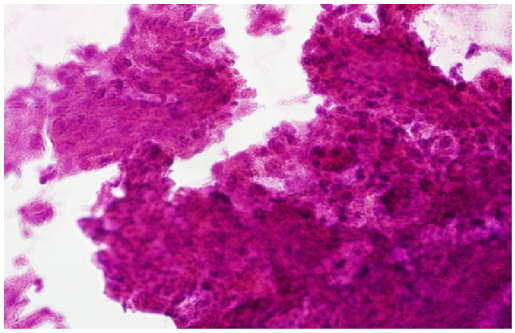


Figure 41: Example of lack of cell differentiation in select G2 samples taken from the JELEN18 dataset.

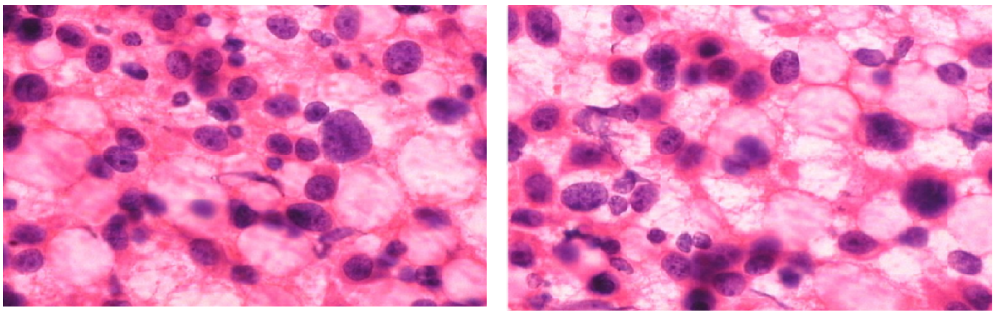


Figure 42: Examples of chromatin variance (smoothly or clumped chromatin) within nuclei in select G2 samples taken from the JELEN16 dataset.

The third challenge was the estimation of the mitosis count (MC) criterion that describes the mitotic activity of cells per high-power field in the diagnosed slide. Generally, each cell divides into two identical cells, then each one of the two produced cells divides to produce two cells, and so on (see

Figure 43). In other words, a mother cell is divided into two identical cells during its life cycle; this division is called bio-mitosis. However, if the mitosis results in more than two identical daughter cells the situation becomes an abnormal case of mitosis which requires the determination of an important malignancy criterion, namely the number of mitosis (see Figure 44). According to Veta *et al.* [114], the proliferative activity of breast tumors, which is routinely estimated (counting the mitotic figures in Hematoxylin and Eosin stained cytology sections), is considered to be one of the most important prognostic markers. However, mitosis counting is exhausting, subjective, and suffers from a low inter-observer agreement limitation.

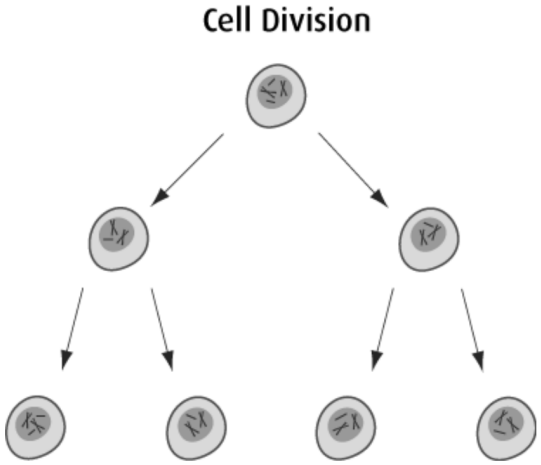


Figure 43: Graphical representation of cell division taken from the [115].

In this study, to estimate the mitosis count malignancy factor we used the methodology of Irshad [116]. The author successfully estimated the MCC in a set of histological images of breast cancer by estimating a set of 143 morphological and textural features for each candidate in different images of red (RGB), blue (RGB), V (HSV), L (Lab) and L (Luv) selected channels. Finally, a decision tree classifier was used.

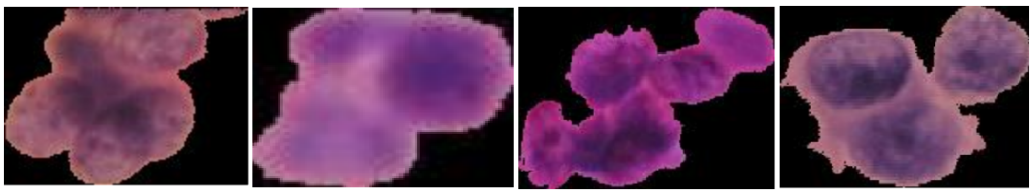


Figure 44: Examples of tumor cells with Mitotic figure activity in select G2 samples taken from the JELEN16 dataset.

Likewise, in order to estimate the MCC, 25 morphological and textural features were calculated from each candidate in the five aforementioned channels. Among these 25 features, 18 features

were used from the previous study and extracted from the GLCM and GLRLM matrices; and seven features (perimeter, uniformity of nuclear size, uniformity of nuclear shape, nuclear irregularity, nuclear smoothness and nuclear roundness) were added based on the experimental results. The goal was to include specific morphological features for this estimation due to the limitations available in the cytological images related to the cell structures. These features were then used to classify the extracted mitosis candidates as mitosis, non-mitosis, and ignored candidates.

In this study, with respect to the modification of MCC for cytological grading schemes, a score of 1 was assigned to cancers in cases with there was 0-1 mitosis per image, a score of 2 in cases with there were 2–4 mitosis, and score of 3 in case with there were >5 mitoses. Thus, based on the described pathology-based characteristics for mitosis counting of CGSs, an SVM with the Gaussian radial basis kernel function was used to estimate the number of mitosis per image for each case. The SVM classifier was trained on a manually-prepared dataset of 1853 candidates that consisted of 129, 203, and 1521 mitosis, non-mitosis and ignored candidates, respectively (see Figure 45). The overall accuracy results of MCC and score classification were 80.43% and 65.22%, respectively, which have been evaluated using 5-fold cross-validation and a Receiver Operating Characteristics (ROC) curve. Thus, a ROC curve was used to evaluate the performance measurement of the classification model of the estimation of MCC. More precisely, in this thesis, we extended the ROC curve (usually used for binary classification problems) to evaluate the multi-class (mitosis, non-mitosis, and ignored candidates) classification for mitosis count by binarizing the output. Thus, we computed a ROC curve and ROC area for each class (see Figure 46). The TP and FP rates were estimated and then used to draw the curves that display the tradeoff between sensitivity and specificity. According to the ROC plot, the closer the curve to the left-hand border and the top border of the ROC space, the higher the accuracy of the test.

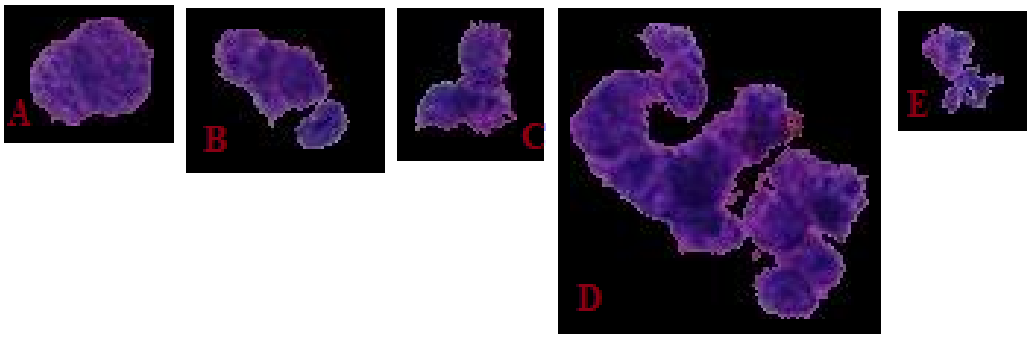


Figure 45: Examples of bio-normal, abnormal mitosis nuclei and ignored candidate of an intermediate malignancy case from JELEN16 dataset. (A) represents a normal mitosis sample, (B-D) represent abnormal mitosis samples and (E) represents an ignored sample.

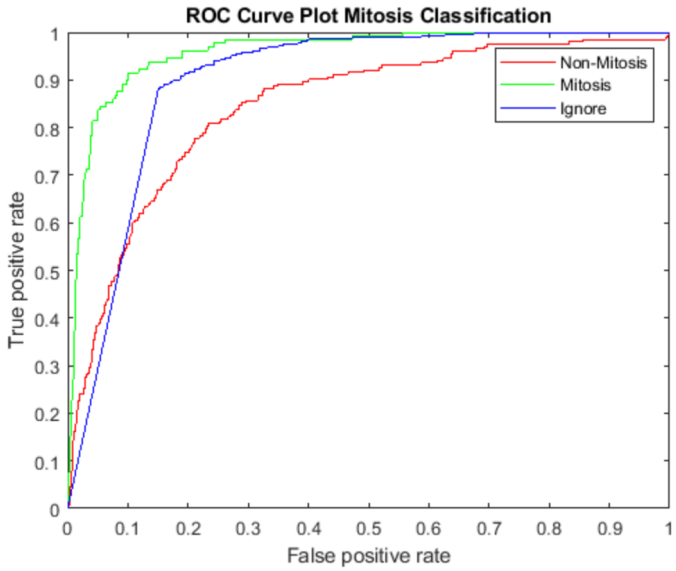


Figure 46: ROC curve results of the three predictors of mitosis, non-mitosis and ignored candidates.

Finally, our last challenge was related to the naked tumor cell nuclei (NTCN) estimation which required some information that could be provided by direct observation (ground truth) such as the standard nuclei size (NS), which is difficult to determine using FNA images. NTCN is a malignancy criterion that interprets the nuclear budding (developing buds) form [41] that is already estimated implicitly by the nuclear polymorphic features and the presence of cell nuclei with highly atypical morphology in the breast tumor. It is characterized by features such as round to polygonal cells, indistinct cell borders, etc. [117]. However, atypical bare nuclei are rarely seen and the clinical significance of this finding is uncertain [118]. Even though some of the estimated pleomorphic and morphological features estimate the naked characteristic which has been defined implicitly, the highly atypical changes in nuclear size were estimated based on taking the ratio of the nuclei number with very large and very small sizes to the total number of nuclei involved in the image to capture the variance range of nuclei size within each image.

According to the implemented cytological schemes (Table 1), features can be computed from a pair of high and low magnification images that belong to each patient. For example, in the implementation of Robinson's [36], Howell's [37], and Mouriquand's [35] grading systems, features from a case of a pair of 100x and 400x magnification images are computed to extract a set of different cellular and nuclear features that are used to classify a given case or patient. For the other three implemented cytological schemes, the information of the 100x images is ignored. Precisely, in the implementation of Taniguchi *et al.* [39], Khan *et al.* [38] and Fisher's[34] grading schemes, a set of different nuclear features are computed from only the 400x magnification images that belong to each case in the

assignment of a malignancy grade to a given case or patient. The final number of the estimated features for each of the six proposed CA-CGS were 87, 89, 88, 88, 89, and 24. They represent Robinson's [36], Fisher's[34], Khan *et al.*'s [38], Taniguchi *et al.*'s [39], Mouriquand's [35], and Howell's [37] grading schemes, respectively. The reason that for Howell's system we estimated only 24 features, is that these 24 features were estimated according to the malignancy characteristics of Howell's CGS. As Howell's system is based on three malignancy criteria - nuclear pleomorphism, mitosis count, and tubule formation - we estimated twenty nuclear pleomorphic features that reflect the change in shape and size of nuclei (nuclear pleomorphism); one for mitosis count; and, three low magnification cellular features that identify the cells' ability to form clusters or disperse within the image (metastasis).

Below is a detailed description of all meaningful estimated sets of features (See Table 1, and the first four paragraphs in this section to get more information about how the estimated features were used efficiently in this thesis to interpret the malignancy characteristics of each of the considered CGS systems). Features extracted from 100x magnification images involve three *cellular structural features* as follows:

- **Area of a group (Ag)**: Cells tend to form clusters within the image. Ag is an average number of nuclei pixels within the cluster. A large value of this feature means a lower number of large clusters in the segmented image (binary image).
- **Number of groups (Ng)**: Represents the number of clusters in the segmented images (binary image). A large value of this feature considers a large number of small groups in the image.
- **Dispersion (D)**: Reflects the spreading of cells within an image and describes the variability among cluster areas (estimated as the average of distance among centroid of all nuclei).

The features of x400 magnification images involve the following sets:

1. The *nuclear pleomorphic features* include the five following features:
    - **Area (A)**: The total number of pixels in a nucleus.
    - **Radius (R)**: Defined as an average of the radial line segments from the centroid of the nucleus to the boundary points.
    - **Eccentricity (Ecc)**: The ratio of the distance between the foci of an ellipse, and the length of its major axis (Ecc = 0 for a circle and Ecc =1 for a line segment).
    - **Convexity (Cx)**: The ratio of the nucleus area and its convex hull, which is the minimum

area of the convex polygon that can contain the nucleus.

- **Compactness (Cn)**: This feature can be estimated by calculating the area of the shape to the shape contour length COMP = Area/Perimeter.

2. The *nuclear morphologic features* involves sixteen features as follows:

   - **Major Axis length (Mxl)**: The length of the major axis of the ellipse (a closed curve composed of points in which the sums of the distances from the foci (two fixed points) to each point are equal).
   - **Minor axis length (Mxl)**: The length of the minor axis of the ellipse.
   - **Mean R, G, and B values**: They represent the mean of pixel values of the nucleus in each of Red, Green and Blue channel, respectively.
   - **Variance R, G, and B values**: They represent the variance of pixel values of the nucleus in each of Red, Green and Blue channels, respectively.
   - **Distance to the Centroid (DTc)**: The distance between the centroid of all of the nuclei in the image to the centroid of a single nucleus.
   - **Mean value of luminance (ML)**: Describes the mean of the perceived brightness of a color.
   - **Variance value of luminance (VL)**: Describes the variance of the perceived brightness of a color.
   - **Uniformity of size(SiU)**: Is the standard deviation of the nucleus area.
   - **Uniformity of shape(ShU)**: Is the standard deviation of the nucleus boundary points.
   - **Irregularity(IRr)**: Uses entropy of intensity values of image pixels as a measure of randomness which is equivalent to irregularity. This, a high value of entropy represents more irregularity and loss represents less irregularity.
   - **Smoothness(Sm)**: Is estimated based on the area solidity (proportion of the pixels in the convex hull that are also in the region) computed as Area/ConvexArea. Thus, a higher solidity value means a smoother shape and a lower value means a less smooth shape.

- **Roundness(Rd)**: Is based on the eccentricity of the boundary points of nuclei as measuring parameter where values near to 0 mean more rounder shapes and values near to 1 mean less round shapes.

3. The *nuclear texture features* involve two sets of features estimated using GLCMs (corresponds to the number of occurrences of pairs of gray levels of an image) and GLRLMs (gives the size of homogeneous runs for each gray level of an image).

   (a) The ten *nuclear co-occurrence features* of GLCMs for horizontal (0°), vertical (90°) and diagonal (45°) & (135°), and eight gray-levels are as follows:

   - **Contrast**: Used to measure the intensity difference between a pixel and its neighbor within an image. Contrast range = $[0 \quad (size(GLCM, 1) - 1)^2]$.

$$Contrast = \sum_{r,c} |r - c|^2 g(r, c) \tag{5}$$

   - **Correlation**: A statistical measure of how correlated a pixel is to its neighbor within an image.

$$Correlation = \sum_{r,c} \frac{(r - \mu_r)(c - \mu_c) g(r, c)}{S_r S_c} \tag{6}$$

   - **Energy**: Also referred to as "unformity" within literature, which means the sum of squared elements in the GLCM. Energy range = [0 1].

$$Energy = \sum_{r,c} g(r, c)^2 \tag{7}$$

   - **Homogeneity**: The closeness of the distribution of elements in the GLCM to the GLCM diagonal. Homogeneity range = [0 1].

$$Homogeneity = \sum_{r,c} \frac{g(r,c)}{1+|r-c|} \tag{8}$$

- **Entropy**: Is a measure of the disorder of a matrix and it is inversely proportional to GLCM energy.

$$Entropy = \sum_{r,c} g(r,c) \log_2 g(r,c) \tag{9}$$

- **Inertia(BR)**: Is defined as the contrast value after converting an RGB image into a blue ratio (BR) image using the following equation:

$$BR = \frac{100 \times B}{1+R+G} \times \frac{256}{1+B+R+G} \tag{10}$$

$$Inertia(BR) = \sum_{r,c} (r,c) g(r,c) \tag{11}$$

- **Cluster-shade(V)**: Is a measure of the skewness of a matrix. When the value of the cluster shade is high, the image is asymmetric. Is estimated base on the mean of the V channel image of HSV color space.

$$Cluster-shade(V) = \sum_{r,c} ((r-\mu_r)+(c-\mu_c))^3 g(r,c) \tag{12}$$

- **Cluster-prominence(V)**: Is also a measure of matrix asymmetry. When the value of cluster prominence is high, the image is less symmetric. We also estimated this feature using the V channel images of HSV color space.

$$Cluster-prominence(V) = \sum_{r,c}((r-\mu_r)+(c-\mu_c))^4 g(r,c) \tag{13}$$

- **Hara-correlation(V)**: Is estimated based on calculating the correlation from the gray-level co-occurrence matrix of the V channel image of HSV color space.

$$Hara-correlation(V) = \sum_{r,c} \frac{(r,c)g(r,c)-\mu^2}{S^2} \tag{14}$$

- **Difference-moment(R)**: Represents the amount of local gray level variation in the red channel image.

$$Difference-moment(R) = \sum_{r,c} \frac{g(r,c)}{1+(r-c)^2} \tag{15}$$

where $g(r,c)$ is the normalized gray-level co-occurrence matrix (its sum is one) for an arbitrary distance & angle, and each element (r,c) in this matrix is the joint probability of occurrence of the pixel pairs with a defined spatial relationship having gray-level values r and c in the image. g_r is the marginal row probabilities & g_c is the marginal column probabilities μ_r and μ_c are the mean gray-level intensities of g_r & g_c, respectively; S_r and S_c are the standard deviations of g_r & g_c, respectively.

(b) The eleven *nuclear run-length features* calculated using GLRLMs for horizontal (0°), vertical (90°), and diagonal (45°) & (135°), and eight gray-levels are as follows:

- **Short run emphasis (SRE)**: used to measure the distribution of short runs.
- **Long run emphasis (LRE)**: used to measure the distribution of long runs.
- **Gray-level non-uniformity (GLN)**: measures the distribution of runs over the gray values.
- **Run length non-uniformity (RLN)**: measures the distribution of runs over the run lengths.

- **Run percentage (RP)**: Measures the ratio between the number of realized runs and the maximum number of potential runs.
- **Low gray-level run emphasis (LGRE)**: measures the distribution of low gray-level values.
- **High gray-level run emphasis (HGRE)**: measures the distribution of the higher gray-level values.
- **Short run low gray-level emphasis (SRLGE)**; emphasizes runs in the upper left quadrant of the GLRLM, where short run lengths and low gray levels are located.
- **Short run high gray-level emphasis (SRHGE)**; emphasizes runs in the lower left quadrant of the GLRLM, where short run lengths and high gray levels are located.
- **Long run low gray-level emphasis (LRLGE)**: emphasizes runs in the upper right quadrant of the GLRLM, where long run lengths and low gray levels are located.
- **Long run high gray-level emphasis (LRHGE)**: emphasizes runs in the lower right quadrant of the GLRLM, where long run lengths and high gray levels are located.

As we mentioned before, the mean and variance were then calculated for each of the mentioned features (pleomorphic, textural and morphologic) giving a total of 84 different nuclear features (see Table 4 and Table 5. Next, an optimal subset of features was selected for classification purpose in the following stage.

In Table 4, the estimated features of the six examples of G3 and G2 cases effectively reflect the biological behavior of cancerous cells, where, for an example, the irregularity-M feature (a measure of randomness), the higher irregularity value indicates the more randomness cells, and in its turn indicates a more advanced malignancy case) values of the high malignancy grade G3 cases are mostly higher than the irregularity-M values of intermediate malignancy grade G2 cases. Another example, the smoothness-V (a measure of area solidity) values for G2 cases are higher than the smoothness-V values of G3 cases which indicate that the lower smoothness case, the higher malignancy grade.

In Table 5, again the estimated three cellular features of the six examples of G3 and G2 cases effectively reflect the biological behavior of cancerous cells. For example, the number-of-groups values of the G3 cases are higher than the number-of-groups values for G2 cases. This is due to the fact that in G3 cases, there are many single nuclei regions that considered as a separated group as compared to G2 cases, where they have individual nuclei and clusters of connected nuclei. Thus, the number of groups in these cases lower than in G3 cases.

Features names	Case1	Case2	Case3	Case4	Case5	Case6
Irregularity-M	0.73939	0.80240	0.76832	0.72529	0.74217	0.69294
Homogeneity-M	0.14583	0.15293	0.15827	0.18716	0.18682	0.21386
Smoothness-V	0.02360	0.01323	0.01724	0.02156	0.02011	0.02759
Roundness-M	0.74948	0.75179	0.75218	0.78747	0.76995	0.82181
Area-M	14.9051	1.16796	1.15148	18.5134	16.7030	2.81764
Radius-M	1.3733	0.5979	0.5937	1.3403	1.2602	0.7787
Eccentricity-V	0.04570	0.02335	0.02011	0.07353	0.07113	0.04486
Entropy-V	0.24242	0.21670	0.24167	0.56932	0.60087	0.51855
Uniformity-of-Size-V	25.2034	8.2784	11.1287	30.0434	28.8085	27.1052
Irregularity-V	0.24242	0.21670	0.24167	0.56932	0.60087	0.51855
Energy-M	0.00315	0.00391	0.00384	0.00628	0.00647	0.00882
Convexity-V	0.0068	0.0002	0.0002	0.0171	0.0159	0.0026
Entropy-V	0.2424	0.2167	0.2417	0.5693	0.6009	0.5185
Mitosis-Cont	1	5	1	8	0	3
Distance-To-Centroid-M	21.8988	18.8125	17.0889	14.9842	16.4705	13.5706
Naked-Nuclei	0.9517	0.9648	0.7742	0.8871	0.9032	0.8269
Mean-Red-M	58.475	82.317	74.375	56.003	58.799	52.897
Pathologist-based grades	G3	G3	G3	G2	G2	G2

Table 4: Some of the calculated high magnification features along with their pathologist-based grades for selected cases of JELEN_MERGE01 dataset.

Features names	Case1	Case2	Case3	Case4	Case5	Case6
Area-of-groups	25141	37387	24395	26610	1902	29509
Number-of-groups	1490	802	1179	88	42	441
Dispersion	432.45	332.74	440.70	369.17	425.23	333.39
Pathologist-based grades	G3	G3	G3	G2	G2	G2

Table 5: The three calculated low magnification features along with their pathologist-based grades for selected cases of JELEN_MERGE01 dataset.

Feature selection stage :

Feature selection plays a central role in helping reduce the high-dimensionality and the noise of the data by removing redundant and irrelevant features. According to the experimental results of this stage in the current study as well as the literature that has employed different methods for feature selection, it is suggested that the presence of ineffective features degrades the performance of the classifiers, particularly if a small training dataset is used to train classification models, and is combined with an additional unbalanced class distribution problem where there are more samples belonging to one class as compared to the other classes (such as in our dataset). Chan *et al.* [119] proposed a computer-aided diagnosis scheme for the classification of true and false detections of masses on a dataset of mammogram images. In the study, the authors employed a stepwise feature selection method and successfully selected a small subset of effective features from huge feature spaces. In alternate steps, one feature is added to or removed from the selected feature set. Based on their method the classifier could correctly identify benign cases from malignant ones using the determined small subset. Goldberg *et al.* [120] selected a subset of features based on the evaluation of the discriminatory ability of the individual feature to distinguish malignant lesions from benign ones in ultrasound images. Wu *et al.* [121] picked a subset of features based on the average value variation of the individual feature between the two classes. In the study of Lo *et al.* [122], priority was assigned to each feature based on the effectiveness of each individual feature on the classification accuracy, priority was assigned for each feature. Then, a subset of features that provided the highest classification accuracy was selected and the features that had the lowest priority were eliminated. Recently, a complete study and survey on the different types (wrappers, embedded, and filter) of feature selection algorithms for classification problems based on the complexity of these algorithms was published by Roffo [123]. The study provided a systematic comparison between these methods

with respect to the accuracy, precision rate, and the stability of the priority of the features with a different dataset. Further, handcrafted features, as well as features extracted with a deep CNN, have been used by the author.

Similarly, in this work, to reduce the high dimensionality of the final feature vector, we examined two supervised learning and filter (correlation-based) feature selection methods including Relieff's and Fisher's feature selection method as proposed by Roffo [123]. The former is an iterative randomized method and evaluates the quality of the features based on how well the values of these features can distinguish examples of data that are similar to each other, while the latter computes the scores of features depending on the ratio of inter-class separation and intra-class variance. According to the obtained classification results using these methods, we selected Fisher's feature selection methods, which showed the more relevant and non-redundant subset of features and yielded better classification results (see Table 6).

Classification stage :

In this section, the classification stage of the implemented cytological schemes is discussed. At this point, the six proposed frameworks based on the six original CGSs have determined feature vectors (see Tables 4 and 5) corresponding to the cytological characteristics for each CGSs, which are then subjected to feature reduction to come up with a 30% optimal subset of features determined experimentally from the total number of features. To determine the malignancy level of FNA biopsies, nine different classifiers that take each feature vector as an input and return one of the two malignancy grades (G2 or G3) as output are used. The classifiers used were LDA, feedfoward neural networks (FFNN), SVM, MLP, DT [62] adaptive boosting of decision trees (DT-AdaBoost) [60] RFDT [64], NB [60] and KNN [62].

The proposed frameworks consist of two classification schemes: case classification and patient classification [103, 52]. The case classification scheme involves a single feature calculation, while patient classification involves multiple feature calculations across multiple cases (different locations at a single tissue). More precisely, in the case classification scheme, a feature vector is independently computed for each pair of images with high and low magnifications, or only the high magnification image that belongs to each patient, and each case is classified separately. The results of the case classification represent multiple classification results for the patient. In the patient classification scheme, to classify a certain patient as G2 or G3, the final classification is achieved by the majority voting of the classification results of the individual cases for that patient. Thus, if 50% or more of the cases belonging to a patient are classified as G2, the patient is classified as G2; otherwise, they are classified as G3.

Features indexes	Features names
1	Irregularity-M
2	Entropy-V
3	Homogeneity-M
4	Smoothness-V
5	Roundness-M
11	Inertia-V
12	Difference-Moment-M
13	Mean-Blue-M
14	Energy-M
15	Luminance-Mean-V
19	Correlation-V
20	Irregularity-V
21	Entropy-M
39	Harra-Gorrelation-V
45	Number-of-Groups
46	Homogeneity-V
59	Uniformity-of-Size-M
60	Harra-Correlation-M
62	Smoothness-M
65	Mean-Red-M
66	Energy-V
67	Eccentricity-V
80	Mean-Green-M
81	Uniformity-of-Shape-M
82	Roundness-V
86	Variance-Blue-M

Table 6: Example of the selected subset of features of the combined high and low magnification feature by Fisher method.

3.1.2 Experimental Results :

In this section, simulation results are presented for the six computer-aided malignancy grading frameworks for FNA biopsies of breast cancer based on the six cytological grading schemes (Table 1). The parameters for nuclei size, feature selection, and the cross-validation value were subjected to different experimental attempts on our datasets with respect to the underlying problem. For nuclei segmentation, the parameter NS, chosen experimentally on the test set of 400x magnification images, was set to 100 pixels, which was determined based on the area that represents the total number of pixels in a nucleus. Given the experimental results on feature vector reduction, for each implemented cytological scheme, the dimensionality of the final feature vector was reduced to 30% of the total number of the estimated features using Fisher's feature selection method as this percentage yielded the highest performance accuracy. The proposed method was tested using the k-fold cross-validation technique for k = 5. More precisely, the dataset was randomly divided into 5 equal-sized subsets (meaning we randomly shuffled the dataset into 5 subsets). Of 5 subsets, one subset was used as the validation data to test the model performance and the other 4 subsets were used as training data. Then, the process was repeated 5 times; each time, one certain subset was used as the validation data. After completing the 5 folds cross-validation, we averaged the 5 folds results to obtain the final estimation of one run.

It is important to mention here that there are inconsistencies in the number of samples belonging to each class (patient and case). In other words, we have different numbers of patients belonging to each class, with very few patients belonging to the malignancy G3 class. Also, each patient has a different number of cases (for example, one patient may have one case, whereas another patient has three or four cases). Moreover, each case has a different number of low and high magnification images. For these limitations, we didn't take patient information into account when dividing cases into training and testing. Although dividing dataset by cases may lead to increased accuracy, there may be an increased bias of the results as the same patient information may appear in both training and testing files. However, cross-validation based on data splitting by patients is the subject of further research and the relevant paper will be submitted for publication in due course.

Further, since using the k-fold cross-validation technique randomly divided the samples into training and testing subsets with each fold, 30 runs for each classifier were performed to optimize the obtained results. Moreover, the 95% confidence intervals were calculated, using the Student's t-distribution, for the obtained classification accuracies to provide a robust indication of the performance abilities of these frameworks.

Using the JELEN_MERGE01 (combined JELEN08 and JELEN16) dataset, Tables 7 and 8 show

the results obtained for the six CA-CGSs. In these tables, for each of the six grading frameworks, the 95% confidence intervals of the accuracies per 30 runs are given for each of the nine classifiers used. According to these results, in terms of case classification, Khan's scheme performs best, followed by Robinson*et al.*'s and then Taniguchi *et al.*'s. In terms of patient classification, Khan's scheme also performs best, followed by Robinson*et al.*'s and then Mouriquand *et al.*'s. Moreover, in terms of classifier results per computer-aided grading scheme, it is readily apparent that the SVM nearly always yields the statistically significant best performance with the three exceptions of Howell's, Khan *et al.*'s and Taniguchi *et al.*'s, where the combination of DT-AdaBoost, FFNN, and RF, respectively, perform marginally better than SVM for case classification. Based on this result, for the remaining experimental results, only the SVM classifier was considered. Thus, with respect to the SVM classifier results, Robinson*et al.*'s CA-CGS achieved the best performance for the case classification problem (consistent with existing pathologist-based literature) while Khan *et al.*'s CA-CGS achieved the best performance for the patient classification problem.

To better understand the performance of CA-CGS based on Robinson's scheme (the best grading system for the case classification problem out of the six proposed CA-CGS based on the SVM classifier results in this study) for case classification, in Table 9, its performance was compared to that of Jeleń *et al.* [83] using the JELEN08 dataset. While, in Table 10, the performance of CA-CGS based on Robinson's scheme was compared to that of Jeleń *et al.* [84] using the JELEN_MERGE01 dataset. Whenever a direct comparison is possible, the current computer-aided framework based on Robinson's scheme outperforms the average accuracy results of Jeleń *et al.* by more than 10% for both studies. These results give significant evidence that cytological images of FNA smears may be lacking the histopathologic features, such as a cellular structure, that a grading scheme like BR is based upon and which, in turn, is the base of the computer-aided malignancy grading system by Jeleń *et al.*.

In terms of classification accuracy evaluation for the proposed CA-CGS systems, the standard overall performance accuracy (estimated by the correct predicted samples to the total number of the predicted samples) has been used as the standard evaluation method for all the proposed CA-CGS systems in this study. Thus, to evaluate the robustness of the new proposed CA-CGSs, we computed the confusion matrix to compare the two classification schemes (case and patient). Also, for all the proposed systems, we computed the average of the accuracy per 30 runs, and the three other measures, that are, the sensitivity, specificity and precision rates based on the SVM classifier results. More precisely, training of the model was performed using 5 cross-validations. Following the training, predictions were made on the test set. The confusion metrics and the accuracy of the test set were calculated according to the prediction of the trained model (see Table 3). This

Cytological Scheme	Classifier	Case Classification	Patient Classification
Robinson's [36]	SVM	**97.37 ± 0.49**%	**95.38 ± 1.15**%
	LDA	96.51 ± 0.54%	93.26 ± 1.04%
	DT+AdaBoost	96.22 ± 0.54%	94.05 ± 1.14%
	RF	95.97 ± 0.59%	94.16 ± 0.98%
	FFNN	95.76 ± 0.96%	92.32 ± 1.49%
	MLP	95.62 ± 1.06%	93.22 ± 1.94%
	DT	93.75 ± 1.04%	92.48 ± 1.39%
	KNN	92.95 ± 1.03%	89.52 ± 1.58%
	NB	83.03 ± 0.39%	83.65 ± 0.71%
Mouriquand's [35]	SVM	**95.12 ± 0.25**%	**95.27 ± 1.24**%
	KNN	94.06 ± 0.90%	88.97 ± 1.27%
	RF	93.56 ± 0.68%	92.32 ± 0.94%
	DT+AdaBoost	93.31 ± 0.24%	92.01 ± 0.21%
	LDA	93.29 ± 0.33%	92.11 ± 1.06%
	MLP	92.40 ± 1.13%	90.21 ± 1.85%
	DT	91.04 ± 1.04%	90.68 ± 1.81%
	FFNN	91.02 ± 0.24%	92.01 ± 0.21%
	NB	83.71 ± 0.08%	82.64 ± 0.44%
Howell's [37]	DT+AdaBoost	**89.93 ± 1.17**%	**89.36 ± 2.02**%
	FFNN	89.05 ± 0.41%	85.60 ± 0.43%
	RF	87.63 ± 1.05%	85.02 ± 1.27%
	KNN	86.88 ± 1.25%	82.91 ± 1.27%
	LDA	86.06 ± 0.98%	84.55 ± 1.12%
	SVM	85.53 ± 0.16%	86.35 ± 0.22%
	DT	83.69 ± 1.48%	83.38 ± 2.22%
	MLP	82.65 ± 0.49%	76.40 ± 0.87%
	NB	81.86 ± 0.25%	80.89 ± 0.21%

Table 7: The 95% confidence interval results of the accuracies of the first three of nine used CA-CGSs using the JELEN_MERGE01 dataset. These three CA-CGSs are based on both low and high magnification features. The bolded values are the best results for each classifier for each CA-CGS.

Cytological Scheme	Classifier	Case Classification	Patient Classification
Fisher's [34]	SVM	**93.44 ± 0.79**%	**92.78 ± 1.34**%
	RF	91.24 ± 0.70%	91.94 ± 1.02%
	DT+AdaBoost	90.82 ± 0.82%	90.84 ± 1.42%
	LAD	90.62 ± 1.03%	90.42 ± 1.26%
	DT	90.28 ± 0.93%	92.27 ± 1.80%
	KNN	88.63 ± 0.84%	84.65 ± 1.54%
	MLP	88.49 ± 1.43%	86.56 ± 2.47%
	FFNN	87.55 ± 1.60%	85.92 ± 2.04%
	NB	76.32 ± 0.90%	76.71 ± 0.89%
Khan *et al.* [38]	FFNN	**97.58 ± 0.62**%	**97.11 ± 0.84**%
	LDA	97.01 ± 0.61%	96.28 ± 1.01%
	SVM	96.75 ± 0.58%	96.24 ± 1.22%
	RF	95.72 ± 0.60%	93.74 ± 1.13%
	DT+AdaBoost	94.81 ± 0.87%	92.32 ± 1.46%
	MLP	93.45 ± 1.40%	92.22 ± 2.46%
	KNN	92.42 ± 1.01%	88.20 ± 1.63%
	DT	92.40 ± 1.39%	92.16 ± 1.65%
	NB	81.82 ± 0.49%	82.16 ± 0.97%
Taniguchi *et al.* [39]	RF	**95.97 ± 0.54**%	**93.80 ± 0.95**%
	LDA	95.61 ± 0.73%	92.91 ± 1.63%
	SVM	95.30 ± 0.79%	94.23 ± 1.18%
	DT+AdaBoost	94.93 ± 0.73%	92.32 ± 1.53%
	MLP	92.99 ± 1.33%	90.79 ± 1.96%
	FFNN	92.85 ± 1.81%	90.68 ± 3.22%
	DT	92.83 ± 1.21%	92.06 ± 1.56%
	KNN	92.12 ± 0.79%	87.14 ± 1.37%
	NB	81.30 ± 0.48%	80.79 ± 1.14%

Table 8: The 95% confidence interval results of the accuracies of the next three of the nine used CA-CGSs using the JELEN_MERGE01 dataset. These three CA-CGSs are based only on high magnification features. The bolded values are the best results for each classifier for each CA-CGS.

	Jeleń *et al.*	Robinson's			
Classification algorithms	**Average accuracy**	**Average accuracy**	**Average sensitivity**	**Average specificity**	**Average precision**
SVM	80.61%	91.52%	70.00%	**100**%	**100**%
MLP	**82.70**%	95.50%	**96.86**%	92.05%	96.86%
LDA	—	96.38%	90.90%	97.74%	90.90%
RF	—	**96.81**%	96.41%	96.97%	92.58%

Table 9: Comparison of case classification based on Robinson's scheme [36] and Jeleń *et al.* [83], using the JELEN08 dataset. Jeleń *et al.* results, where available were taken from [83]. Best results indicated in bold.

	Jeleń *et al.*			Robinson's		
Classification algorithms	**Average accuracy**	**Average sensitivity**	**Average specificity**	**Average accuracy**	**Average sensitivity**	**Average specificity**
SVM	76.24%	91.23%	56.82%	**97.59**%	94.04%	98.47%
MLP	76.2%	68.4%	**86.4**%	95.83%	97.82%	87.57%
FFNN	**87.1**%	**94.7**%	77.3%	94.63%	**98.52**%	78.99%
LDA	—	—	—	96.38%	90.90%	97.74%
RF	—	—	—	95.78%	84.14%	**98.67**%

Table 10: Comparison of case classification based on Robinson's scheme [36] and Jeleń *et al.* [84], using the JELEN_MERGE01 dataset. Jeleń *et al.* results, where available were taken from [84]. Best results indicated in bold.

process is repeated 30 times to get 30 accuracies. Then, the average accuracy was estimated for 30 runs to achieve the optimized model result. Further, the overall confusion matrices of 30 runs were computed to evaluate and describe the performance of the proposed systems on the two classification schemes (case and patient), and to investigate the misclassified cases and patients within these 30 runs with respect to the malignancy grading problem, as shown in Figure 47, where the overall confusion matrices of 30 runs were calculated for Robinson's and Khan *et al.*'s systems, respectively.

Due to the fact that classification accuracy alone is not enough to evaluate the robustness of classification model with respect to solving the problem at hand, specifically when used with an imbalanced dataset, as we mentioned in the previous paragraph, we computed three measures from binary classification where we considered the G2 classification as the negative class and the G3 classification as the positive class. These measures are i) sensitivity—the proportion of the correct G3 classifications over all G3 classifications; specificity—the proportion of the correct G2 classifications over all G2 classifications; and precision—the proportion of the correct G3 classifications over all correct G3 and G2 classifications (see Tables 11 and 12).

In Figure 47, the diagonal cells to right of the matrices contain the number of the correctly (TN and TP) classified samples by Robinson's and Khan's grading systems while the diagonal cells to left cells of the matrices contain the number of the incorrectly (FP and FN) classified samples. For instance, in the matrix (a), for the case confusion problem, 3922 (78.8%) cases out of the 4980 cases were correctly classified as malignancy grade MG2. Similarly, 921 (18.5%) cases were correctly classified as malignancy grade MG3. Also, 69 (1.4%) of the MG3 cases were incorrectly classified as MG2 and 68 (1.4%) of the MG2 cases were incorrectly classified as MG3. A similar description can be used to illustrate the patient confusion matrix (b) of Robinson's system, and the confusion matrices (c and d) of Khan *et al.*'s system results.

In Table 11, for case classification, an extension of the comparison between Robinson's and the other computer-aided classification schemes was performed using the SVM classifier on the JELEN_MERGE01 dataset. It was shown that beyond the best average accuracy and sensitivity, Robinson's also has a very high specificity rate of 98.47% which is very close to the highest specificity of 98.64%, obtained by Khan *et al.*'s system. Regarding the precision, Robinson's yields a relatively low result of 93.84%, being well outperformed again by Khan *et al.*'s which recorded 94.24%.

In Table 12, for patient classification, the comparison between Khan *et al.*'s and the other computer-aided classification schemes was extended using the SVM classifier on the JELEN_MERGE01

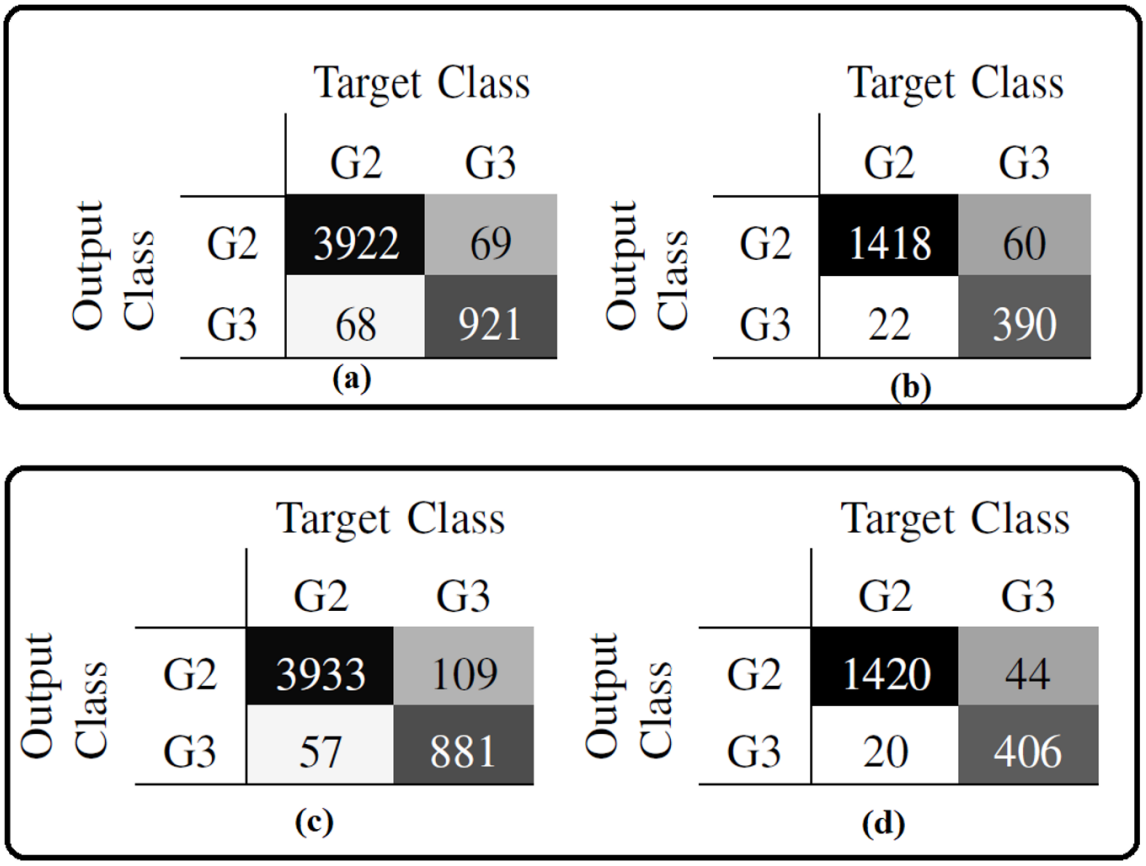


Figure 47: Example of the calculated confusion matrices for 30 runs, (a) and (c) Case classification confusion matrix, (b) and (d) Patient classification confusion matrix of Robinson's and Khan *et al.*'s systems, respectively, using JELEN_MERGE01 dataset.

Cytological schemes	Average Accuracy	Average sensitivity	Average specificity	Average precision
Robinson's ☆	**97.57**%	**93.93**%	98.47%	93.84%
Fisher's	93.35%	75.25%	97.84%	89.68%
Khan *et al.* *	96.66%	88.68%	**98.64**%	**94.24**%
Taniguchi *et al.*	95.30%	84.94%	97.87%	90.92%
Mouriquand's	95.56%	88.68%	97.26%	89.09%
Howell's	85.64%	66.16%	90.47%	63.48%

Table 11: Evaluation results on case classification using the SVM classifier on the JELEN_MERGE01 dataset for all the six cytological grading frameworks. Best results indicated in bold. ☆ - The first best CA-CGS. * - The second best CA-CGS.

dataset. It was shown that beyond the average accuracy, Khan *et al.*'s also has the best sensitivity result. For specificity, Khan *et al.*'s has a respectable result of 98.61%, although Fisher's achieved the best rate at 99.79%. For precision, Khan *et al.*'s has a relatively low result of 94.66% being well outperformed again by Fisher's at 99.20%.

Cytological schemes	Average accuracy	Average sensitivity	Average specificity	Average precision
Robinson's *	95.29%	85.77%	98.26%	93.99%
Fisher's	92.69%	70.00%	**99.79**%	**99.20**%
Khan *et al.* ☆	**96.24**%	**89.33**%	98.40%	94.66%
Taniguchi *et al.*'s	94.23%	80.66%	98.47%	94.47%
Mouriquand's	95.02%	80.44%	99.58%	98.41%
Howell's	86.08%	61.33%	93.81%	75.91%

Table 12: Evaluation results on patient classification using the SVM classifier on the JELEN_MERGE01 dataset for the six cytological grading frameworks. Best results indicated in bold. ☆ - The first best CA-CGS. * - The second best CA-CGS.

According to the achieved results of the six original CA-CGSs proposed in this section, the best

results were obtained from the SVM classifier for the computer-aided version of Robinson's and Khan *et al.*'s cytological grading systems with accuracies of 97.57% and 96.66% for case classification (where a case is a pair of 100x and 400x magnification images for a patient) and 95.29% and 96.24% for patient classification, respectively.

However, a shortcoming of the six CA-CGSs is the discriminatory power of low magnification features which were missing in three of those six CA-CGSs as well as the low performance on the minority class (G3 Classification) prediction as compared to majority class (G2 classification) prediction. Therefore, one aspect of the thesis is to examine the influence of the low magnification features (100x) on the accuracy performance of the CA-CGSs. Thus, as part of this examination, in the next section, three modified CA-CGSs are proposed and discussed by adding low magnification features to Fisher's, Khan *et al.*'s and Taniguchi *et al.*'s cytological grading systems which, until now, considered only high magnification features (400x).

3.2 Developing three computer-aided cytological grading systems by modifying the three high magnification feature-based cytological grading systems

As mentioned in the previous section, six computer-aided cytological grading systems (CA-CGSs) were proposed for FNAs of breast cancer based on six published cytological malignancy grading schemes (CGSs) used by pathologists. In these systems, to maintain a strong connection between the decision-making process of the pathologist using one of these CGSs and the determination of the computer-aided CGSs, the features used in our classification systems were carefully tailored to best express the cytological characteristics used by the pathologist. In this section, due to the importance of the low magnification features that were examined in the previously proposed CA-CGSs, three new CA-CGSs were proposed by modifying three of the CA-CGSs that are proposed based on the high magnification features only. To do so, we will add three low magnification features extracted from the low magnification images to the three aforementioned CA-CGSs with the aim of testing the impact of these cellular features on the overall accuracy performance. Three additional classification frameworks were tested based on these modified CA-CGSs.

3.2.1 The methodology

To modify the three considered CA-CGSs, all the fundamental stages of the proposed methodology in the previous section will be the same for the three modified CA-CGSs in this section. One change

is done in the classification stage where the features can now be computed from a pair of high and low magnification images that belong to each case instead of using only the high magnification features as used in the original versions. Therefore, features from a case of a pair of 100x and 400x magnification images are computed to extract a set of different cellular and nuclear features that are used later to classify a given case or patient. The final number of the estimated features for these three modified CA-CGSs were 92, 91, and 91. They represent the modified Fisher's, modified Khan *et al.*'s and modified Taniguchi *et al.*'s grading schemes, respectively.

3.2.2 Experimental results

To evaluate the performance of the three modified CA-CGSs, this section discusses several comparisons among all nine proposed CA-CGSs as well as between the modified systems and their original versions in terms of case and patient classification results. Further, the impact of the added low magnification features on boosting the performance of the overall systems' accuracies were examined. To begin with, in Table 13, we display the results obtained from the three modified CA-CGSs. For the case classification, the SVM classifier gave the best results out of the nine classifiers used for each of the three modified CA-CGSs. While, for patient classification, the SVM classifier also showed the best accuracy performance with one exception where the DT-AdaBoost classifier achieved the best accuracy performance for the computer-aided version of the modified Khan *et al.*'s system.

To better understand the performance of the modified Khan *et al.*'s scheme (the best system out of the nine proposed CA-CGSs for case classification problem based on SVM classifier) for case classification, Table 14 compares the performance of the modified Khan *et al.*'s and Jeleń *et al.* [83], using the JELEN08 dataset. Table 15 compares the performance of the modified Khan *et al.*'s and Jeleń *et al.* [84] using the JELEN_MERGE01 dataset. The current computer-aided framework based on the modified Khan *et al.*'s scheme outperforms the average accuracy results of Jeleń *et al.* by about 15% in [83] and about 7% in [84].

To evaluate the robustness of the three modified CA-CGSs, the average accuracy, sensitivity, specificity, and precision were estimated for 30 runs to get the optimized model's result. Further, the overall confusion matrices of 30 runs were computed to evaluate the performance of the three modified frameworks on the case and patient classification schemes.

Two additional comparisons were performed in this study. The first comparison was performed in terms of the case and patient classification results among all nine CA-CGSs after adding the low magnification features to the aforementioned three systems, while the second comparison was performed in terms of case and patient classification results among the originally proposed versions

Cytological Scheme	Classifier	Case Classification	Patient Classification
Modified Fisher's	SVM	**96.45 ± 0.76%**	**95.71 ± 0.8%**
	RF	92.78 ± 0.79%	92.15 ± 0.10%
	DT+AdaBoost	93.06 ± 0.9%	92.64 ± 0.8%
	LAD	91.37 ± 0.8%	89.10 ± 0.11%
	MLP	90.84 ± 1.17%	88.62 ± 1.90%
	FFNN	90.58 ± 1.57%	89.04 ± 2.10%
	DT	90.40 ± 0.92%	90.95 ± 0.99%
	KNN	89.91 ± 1.23%	87.77 ± 2.09%
	NB	78.81 ± 0.86%	80.00 ± 1.31%
Modified Khan *et al.*'s	SVM	**97.77 ± 0.57%**	**96.50 ± 1.14%**
	DT+AdaBoost	97.52 ± 1.62%	98.22 ± 1.93%
	LDA	97.44 ± 0.67%	95.76 ± 1.18%
	MLP	95.94 ± 0.93%	93.65 ± 1.43%
	RF	95.66 ± 0.76%	93.28 ± 1.23%
	DT	93.73 ± 1.32%	92, 96 ± 1.78%
	FFNN	95.48 ± 1.13%	92.85 ± 1.78%
	KNN	92.87 ± 1.09%	90.10 ± 1.66%
	NB	83.33 ± 0.49%	84.97 ± 1.02%
Modified Taniguchi *et al.*'s	SVM	**97.24 ± 0.69%**	**95.29 ± 1.19%**
	LDA	96.62 ± 0.70%	93.49 ± 1.26%
	FFNN	96.32 ± 1.18%	92.75 ± 1.61%
	DT+AdaBoost	96.14 ± 0.64%	93.65 ± 0.99%
	MLP	96.14 ± 0.84%	93.70 ± 1.54%
	RF	95.74 ± 0.53%	93.33 ± 0.90%
	DT	93.55 ± 1.24%	92.48 ± 1.78%
	KNN	92.89 ± 1.05%	89.57 ± 1.45%
	NB	82.85 ± 0.31%	83.12 ± 0.73%

Table 13: The 95% confidence interval results of the accuracies of the three modified CA-CGSs using the JELEN_MERGE01 dataset. These three CA-CGSs are the same CGSs from Table 31 that have been modified by the addition of low magnification features. The bolded values are the best results for each classifier for each CA-CGS.

	Jeleń *et al.*	Modified Khan *et al.*			
Classification algorithms	Average **accuracy**	**Average accuracy**	**Average sensitivity**	**Average specificity**	**Average precision**
SVM	80.61%	**97.53**%	92.56%	99.49%	98.64%
MLP	**82.70**%	95.21%	**97.87**%	88.46%	95.61%
FFNN	—	92.02%	93.33%	88.71%	95.49%
DT+AdaBoost	—	91.37%	68.71%	**100**%	**100**%
LDA	—	91.66%	81.79%	95.55%	88.34%

Table 14: Comparison of case classification based on modified Khan *et al.*'s scheme [38] and Jeleń *et al.* [83], using the JELEN08 dataset. Jeleń *et al.* results, where available were taken from [83]. Best results indicated in bold.

	Jeleń *et al.*			Modified Khan *et al.*		
Classification algorithms	**Average accuracy**	**Average sensitivity**	**Average specificity**	**Average accuracy**	**Average sensitivity**	**Average specificity**
SVM	76.24%	91.23%	56.82%	**97.77**%	95.65%	98.29%
MLP	76.2%	68.4%	**86.4**%	95.94%	**98.62**%	85.15%
FFNN	**87.1**%	**94.7**%	77.3%	95.48%	98.34%	83.93%
DT+AdaBoost	—	—	—	97.52%	95.96%	99.09%
LDA	—	—	—	97.45%	90.70%	**99.12**%

Table 15: Comparison of case classification based on modified Khan *et al.*'s scheme [38] and Jeleń *et al.* [84], using the JELEN_MERGE01 dataset. Jeleń *et al.* results, where available were taken from [84]. Best results indicated in bold.

of Fisher's, Khan *et al.*'s and Taniguchi *et al.*'s grading systems and their modified versions.

Starting with the first comparison, in Table 16, for case classification, the comparison between modified Khan *et al.*'s and the other computer-aided classification schemes using SVM on the JELEN_MERGE01 dataset is presented. We can see that beyond average accuracy and sensitivity, the modified Khan *et al.*'s also achieved the same high specificity of 98.29% like the original Khan *et al.*'s system. For precision, the modified Khan *et al.*'s achieved a relatively low precision of 93.34%. However, this rate is very close to the best precision of 94.24%, obtained by Khan *et al.*'s system.

Cytological schemes	**Average accuracy**	**Average sensitivity**	**Average specificity**	**Average precision**
Robinson's *	97.28%	93.33%	98.27%	93.09%
Fisher's	93.35%	75.25%	97.84%	89.68%
Khan *et al.*'s	96.66%	88.68%	**98.64**%	**94.24**%
Taniguchi *et al.*'s	95.30%	84.94%	97.87%	90.92%
Mouriquand's	95.56%	88.68%	97.26%	89.09%
Howell's	85.64%	66.16%	90.47%	63.48%
Modified Fisher's	96.44%	88.89%	98.32%	92.93%
Modified Khan *et al.*'s ☆	**97.77**%	**95.65**%	98.29%	93.34%
Modified Taniguchi *et al.*'s	97.24%	93.03%	98.29%	93.16%

Table 16: Evaluation results on case classification using the SVM classifier on the imbalanced JELEN_MERGE01 dataset for all nine CA-CGSs after adding the low magnification fetuses. Best results indicated in bold. ☆ - The first best CGS. * - The second best CGS.

In Table 17, for patient classification, the comparison between modified Khan *et al.*'s and the other computer-aided classification schemes using the SVM classifier was expanded on the JELEN_MERGE01 dataset. We can see that, beyond average accuracy and sensitivity, the modified Khan *et al.*'s also has a respectable rate of specificity at 97.98% though Fisher's achieved 100% of specificity. For precision, modified Khan *et al.*'s has a relatively low precision of 93.50% being well outperformed again by Fisher's at 100%.

The second comparison was performed on Fisher's, Khan *et al.*'s and Taniguchi *et al.*'s systems

Cytological schemes	**Average accuracy**	**Average sensitivity**	**Average specificity**	**Average precision**
Robinson's	95.29%	85.77%	98.26%	93.99%
Fisher's	92.69%	70.00%	99.79%	99.20%
Khan *et al.*	96.24%	89.33%	98.40%	94.66%
Taniguchi *et al.*'s	94.23%	80.66%	98.47%	94.47%
Mouriquand's	95.02%	80.44%	99.58%	98.41%
Howell's	86.08%	61.33%	93.81%	75.91%
Modified Fisher's *	95.71%	82.00%	**100**%	**100**%
Modified Khan *et al.*'s ☆	**96.50**%	**91.77**%	97.98%	93.50%
Modified Taniguchi *et al.*'s	95.29%	86.22%	98.12%	93.60%

Table 17: Evaluation results on patient classification using the SVM classifier on the JELEN_MERGE01 dataset for all nine CA-CGSs after adding the low magnification features. Best results indicated in bold. ☆ - The first best CA-CGS. * - The second best CA-CGS.

and the modified versions of these CA-CGSs based on the SVM classifier results, considering the JELEN_MERGE01 dataset. In Table 18, for case classification, the three modified computer-aided frameworks of Fisher's, Khan *et al.*'s and Taniguchi *et al.* 's outperform the original versions of these systems by 3.9%, 1.11% and 1.94%, respectively. For instance, this means that modified Fisher's system will be able to correctly classify 4 more cases in a group of 100 cases.

In Table 19, for patient classification, the computer-aided frameworks based on the modified Fisher's, Khan *et al.*'s, and Taniguchi *et al.*'s systems outperform their original versions by 3.2%, 0.26% and 1.6%, respectively. Consequently, this implies that Fisher's scheme will be able to correctly classify 2 more patients, assuming the 63 patients presented in the study. Moreover, the computer-aided framework based on the modified Khan *et al.*'s system performers best compared to the other eight cytological grading systems for both case and patient classification tasks.

Cytological schemes	**Average accuracy**	**Average sensitivity**	**Average specificity**	**Average precision**
Original CA-CGSs				
Fisher's	93.35%	75.25%	97.84%	89.68%
Khan *et al.*	96.66%	88.68%	**98.64**%	**94.24**%
Taniguchi *et al.*	95.30%	84.94%	97.87%	90.92%
Modified CA-CGSs				
Modified Fisher's ☆	96.44%	88.89%	98.32%	92.93%
Modified Khan *et al.*'s	**97.77**%	**95.65**%	98.29%	93.34%
Modified Taniguchi *et al.*'s	97.24%	93.03%	98.29%	93.16%

Table 18: Evaluation results on case classification using the SVM classifier on the JELEN_MERGE01 dataset for the three original and three modified CA-CGSs. Best results indicated in bold. ☆ - The best improved result after adding 100x features.

According to the achieved results after modifying the Fisher's, Khan *et al.*'s, and Taniguchi *et al.*'s systems, the best results were obtained from the SVM classifier for the computer-aided version of the modified Khan et al.'s with accuracies of 97.77% and 96.50% for the case and patient classification, respectively. Further, with respect to all nine CA-CGSs, the best two accuracies for the case classification were obtained by the modified Khan *et al.*'s and Robinson's with 97.77% and 97.28%,

Cytological schemes	Average accuracy	Average sensitivity	Average specificity	Average precision
Original CA-CGSs				
Fisher's	92.69%	70.00%	99.79%	99.20%
Khan *et al.*	96.24%	89.33%	98.40%	94.66%
Taniguchi *et al.*'s	94.23%	80.66%	98.47%	94.47%
Modified CA-CGSs				
Modified Fisher's ☆	95.71%	82.00%	**100%**	**100%**
Modified Khan *et al.*'s	**96.50%**	**91.77%**	97.98%	93.50%
Modified Taniguchi *et al.*'s	95.29%	86.22%	98.12%	93.60%

Table 19: Evaluation results on patient classification using the SVM classifier on the JELEN_MERGE01 dataset for the three original GGSs and their modified versions. Best results indicated in bold. ☆ - The best improved result after adding 100x features.

respectively. Meanwhile, for the patient classification, the best two accuracies of 96.50% and 95.71% were obtained by the modified Khan *et al.*'s and modified Fisher's, respectively.

3.3 Conclusions

The discriminatory power of the handcrafted extracted features on the accurate determination of the malignancy grades for FNA biopsies of breast cancer was studied in the first section of this chapter. Furthermore, the impact of the low magnification features on boosting the performance accuracy of the malignancy grading systems of breast cancer was investigated in the second section.

To achieve the mentioned goals, six computer-aided cytological grading systems (CGSs) for FNA biopsies of breast cancer based on the six published cytological malignancy grading schemes used by pathologists (Robinson's CGS; Khan *et al.*'s CGS; Fisher's CGS; etc.) were proposed in section 3.1. In section 3.2, three modified CGSs were proposed by adding the low magnification features that were estimated from the low magnification images to each of the Fisher's, Khan *et al.*'s and Taniguchi *et al.*'s systems. The idea behind the modification of these systems was to add low magnification cellular features that reflect the tubule formation information missing in the cytological images.

For most classifiers used, for all nine CA-CGSs, the achieved average accuracy reached 90% for any given sample (for both case and patient classification) with one exception where Howell's system achieved 80% to 85% average accuracy for both the case and patient classification. Considering the results obtained from only the original six CA-CGSs and based on the SVM classifier, Robinson's system achieved the highest accuracy performance of 97.57% out of the proposed six CA-CGSs for the case classification which is consistent with pathologist-based studies. Khan *et al.*'s system achieved the highest accuracy performance of 96.24% out of the six original CA-CGSs for patient classification.

On the other hand, considering all nine CA-CGSs, the best results, for case classification, were obtained by the SVM classifier for computer-aided versions of the modified Khan *et al.*'s and Robinson's systems with accuracies of 97.77% and 97.28%, respectively. For patient classification, the best results were obtained by the SVM classifiers for computer-aided versions of the modified Khan *et al.*'s and modified Fisher's systems with accuracies of 96.50% and 95.71%, respectively. Furthermore, the modified Khan *et al.*'s achieved the highest accuracy performance for both case and patient classification problems out of all nine proposed CA-CGSs. Finally, the overall performance accuracies of the three modified CA-CGSs outperformed their respective original versions. In this chapter, the discriminatory power of the handcrafted extracted features was addressed to effectively determine the malignancy grades for FNA biopsies of breast cancer. On the other hand, the impact of the low magnification features on boosting the performance accuracy of the malignancy grading systems of breast cancer was studied. Medical image classification based on traditional machine learning algorithms requires a thorough analysis of image components using image processing algorithms. In this chapter, the focus was to adopt different algorithms to precisely segment the regions of interest in the FNA biopsy images. Further, different methods were used to handcraft meaningful and powerful features that simulate the pathology-based criteria for the cytological grading of breast cancer cells. To achieve this goal, there was an accuracy versus complexity trade-off in the above-mentioned simulations.

Although the results were satisfactory, by closely analyzing the sensitivity of Fisher's and Howell's systems, which were merely 70% and 60%, respectively, it is apparent that the grade G3 classification rate was not promising even when the precision was very high. This problem is due to the class imbalance which means that there is more G2 data compared to G3 and consequently, the model was biased for G2 which leads to the misclassification of G3. Therefore, the next chapter considers the problem of class imbalance.

Chapter 4

Data sampling techniques to handle imbalance classification for malignancy grading of breast cancer

The performance of existing traditional learning algorithms can be affected by different learning aspects. Numerous studies in the fields of machine learning and data mining have noted that these aspects may include a class imbalance or unbalanced data problems in classification models [89]. When a dataset contains more samples belonging to one class (majority class) as compared to other classes (minority classes), it is called an imbalanced dataset. In class imbalanced classification, classifiers can achieve very poor performance accuracies on the minority classes in comparison to the majority class. This poor performance with regards to the minority classes is undesirable. In the case of unbalanced data, in general, traditional classification algorithms tend to focus on minimizing the misclassification on predicted classes. These classifiers are likely to ignore the difference between minority and majority classes which leads to a decrease in the sensitivity of the classification system. In this chapter, the focus is on the performance improvement of the nine CGSs proposed in the previous chapter when used with an imbalanced class distribution dataset, resulting in imbalanced classification problems. To improve the performance of these systems when dealing with imbalanced data, we consider the use of two data sampling techniques, namely oversampling the minority class and undersampling the majority class, as well as the Hybrid RUSBoost ensemble-learning algorithm that combines random undersampling and boosting techniques to adjust the data class distribution.

4.1 The methodology

The main objective of using data balancing techniques is to enhance the performance accuracy of the nine cytological grading systems proposed in the previous chapter, which face difficulties related to the nature of the data used (imbalanced dataset). Thus, with the purpose of alleviating the influence of the imbalanced class distribution problem on the classification accuracy, two data sampling and hybrid ensemble-boosting approaches are performed to adjust the class distribution by equally re-balancing the two classes (50:50) before the classification stage. The data sampling techniques include the use of oversampling or undersampling to achieve a balanced (50:50) class distribution. The oversampling technique overcomes the imbalanced class distribution by adding samples to the minority class by either duplicating the samples or adding new samples, whereas the undersampling technique handles this problem by eliminating samples from the majority class. In this study, the oversampling technique was performed by randomly selecting and duplicating some samples from the malignancy grade G3 (the minority class)to balance the minority and majority classes, while the undersampling technique was performed by randomly selecting and deleting some samples from the malignancy grade G2 (the majority class) to balance the minority and majority classes. On the other hand, the RUSBoost boosting-based ensemble learning approach [124] combines boosting (using the AdaBoost algorithm) with random undersampling to create an ensemble classification model (used to make predictions on new data using a collection of individual learners such as DT, KNN, and discriminant). This approach improves the performance accuracy of models trained on skewed data. When RUSBoost is employed, it will remove samples from the majority class until minority and majority class samples are equal in number. AdaBoost, a boosting algorithm, iteratively modifies sample weights that were incorrectly classified in order to ensure correct classification in subsequent iterations. This technique is particularly useful as minority class samples are likely to be misclassified when they are assigned higher weights in different iterations. When combined with RUSBoost, which uses random undersampling, the class distribution is rebalanced and weak learners are improved. For the RUSBoost method, two hyperparameters were selected experimentally for an optimal combination of speed and accuracy. To begin with, we specified the ratio of undersampling of the majority class with respect to minority class as [2 1]. Thus, since we have 33 G3 samples and 133 G2 samples, by setting this ratio every learner in the ensemble was trained on 66 samples of G2 class and 33 samples of G3 class. Further, we set the number of ensemble learning cycles as 300 trees to be performed. Typically, a good prediction model requires between a few hundred to a few thousand weak learners. Thus, after different values of iterations were examined, such as 100, 200, 300, 400, 500, and 600, we selected 300 as the number of iterations for an optimal combination of speed and accuracy.

After applying the data rebalancing technique, the newly obtained results from the nine CA-CGSs after adjusting the class distribution in the dataset were compared with their old results using the imbalanced dataset obtained in the previous chapter. The comparison was done in terms of the overall performance accuracies and the other important measurements to evaluate the performance of each classifier.

4.2 Experimental results

The methodology of the CA-CGSs remains the same after adjusting the class distribution for these systems. Thus, for nuclei segmentation, the parameter NS, chosen experimentally on the test set of 400x magnification images, was set to 100 pixels. For the feature selection process, for each implemented cytological scheme, using Fisher's feature selection method, the final feature vectors were reduced by 70%. In addition, the dataset was divided randomly into training and testing (validation) subsets using the k-fold cross-validation technique for $k = 5$. Once again, we didn't take patient information into account when dividing cases into training and testing due to the limitations associated with the datasets (see Experimental Results in Chapter 3). Moreover, to optimize the accuracy performance of the results, 30 runs for each classifier were performed. Lastly, the 95% confidence intervals were calculated, using the Student's t-distribution, for the obtained classification accuracies to provide a robust indication of the performance abilities of these frameworks.

According to the results obtained in chapter 3, the best classification results were obtained for the SVM classifier using the JELEN_MERGE01 dataset, considering only the original six CA-CGSs results, for the case classification, Robinson's and Khan *et al.*'s systems achieved the best accuracies which were of 97.57% and 96.24%, respectively. For patient classification, Khan *et al.*'s and Robinson's systems achieved the highest accuracies of 9.24% and 95.29%, respectively. Whereas, considering the results obtained of all nine CA-CGSs, the modified Khan *et al.*'s and Robinson's systems achieved the highest accuracies of 97.77% and 97.28%, respectively, for case classification, while the modified Khan *et al.*'s and modified Fisher's systems achieved the highest accuracies of 96.50% and 95.71%, respectively, for patient classification.

In this chapter, to eliminate the imbalanced class distribution problem in the used JELEN_MERGE01 dataset, as well as to enhance the overall performance accuracy of the proposed nine CA-CGSs and specifically, to improve the sensitivity rate of some of these proposed CA-CGSs, two data resampling and RUSBoost approaches were used. The JELEN_MERGE01 dataset consists of 266 samples, where 33 of the samples belong to the G3 (positive class), and 133 samples belong to the G2 (negative class). After applying the mentioned rebalance data techniques to adjust the class distribution

to a 50:50 ratio among the two classes, the new training dataset using the oversampling technique was contained 266 samples, while the new training dataset using the undersampling technique was contained 66 samples. With the oversampling technique, there is no loss in information. However, the training time will be increased due to the added samples. Also, it can lead to an overfitting problem, because the classification algorithm learns to classify the same samples multiple times. In contrast, with the undersampling technique, there is a loss of information due to the deleted samples, but the training time of the classifiers will be decreased. On the other hand, though the random undersampling technique in the RUSBoost approach leads to loss of information, the boosting algorithm implicitly handles this problem [124]. The only side effect that we noticed with this algorithm is that it necessitates long training times, required to train the boosting-ensemble as compared with other data sampling techniques. According to the obtained results in this thesis, the oversampling technique achieved the best overall results for the used dataset. As a comparison between the two data sampling techniques and the RUSBoost approach, the experimental results, shown in Figures 48–51, demonstrate that, among the nine considered classifiers, overall the classifiers considered, data sampling techniques outperformed the average accuracy of the RUSBoost approach by at least 8% for all the original six CA-CGSs (with 95% confidence intervals indicated) for the case and patient classification.

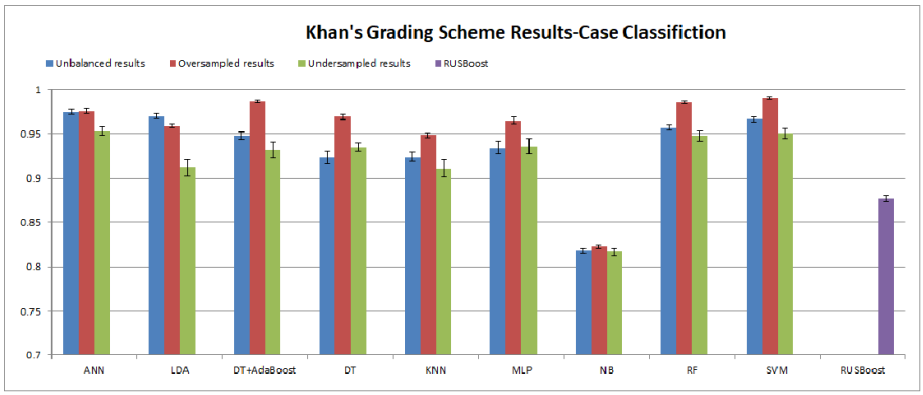


Figure 48: Khan's case classification results before and after adjusting the class distribution for all the used classifiers.

To evaluate the robustness of the proposed frameworks on the test subset using the rebalanced dataset, confusion matrices were used to compare the two classification schemes (case and patient schemes). Further, the accuracy average was computed as well as the three other measures: the sensitivity, specificity, and precision rates as shown in the below Tables 20–29.

On the other hand, to evaluate the best classification results (best classifier), for case classification, considering only the six original CGSs, in Tables 20 and 21, the comparison between the best two

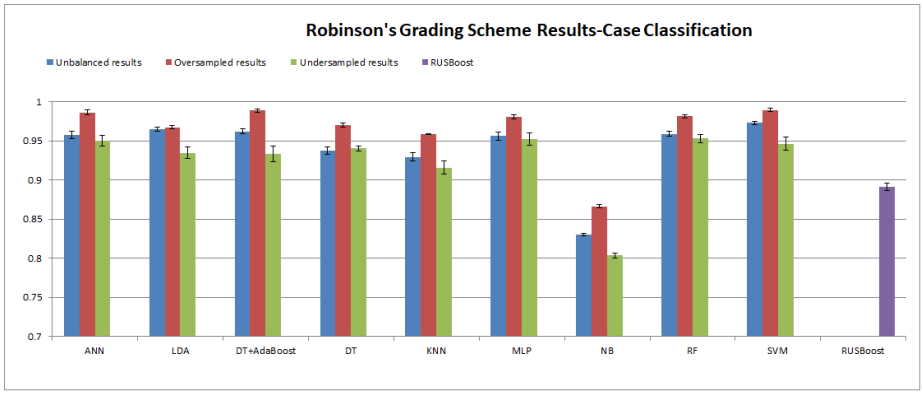


Figure 49: Robinson's case classification results before and after adjusting the class distribution for all the used classifiers.

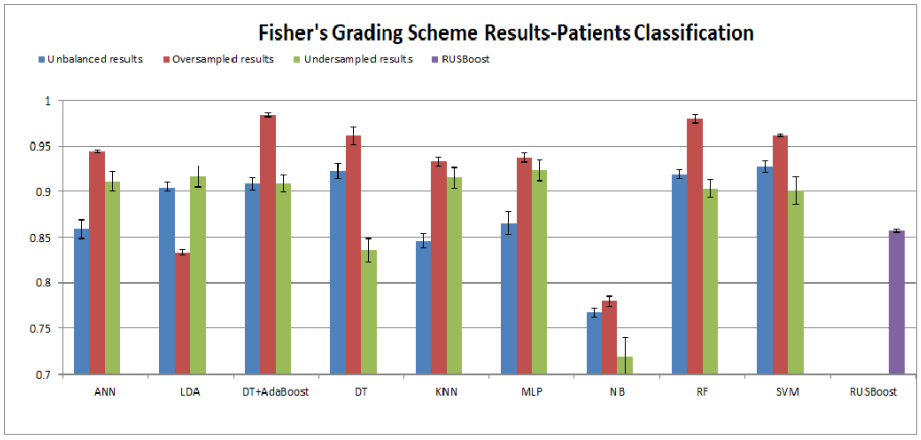


Figure 50: Fisher's patient classification results before and after adjusting the class distribution for all the used classifiers.

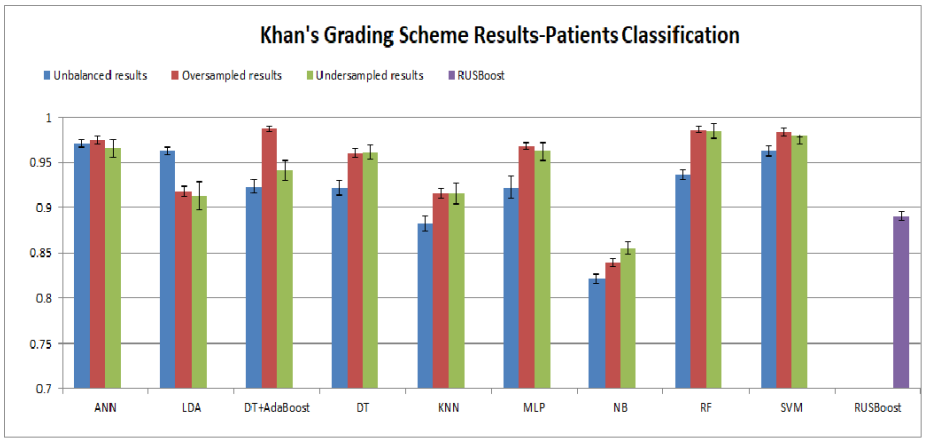


Figure 51: Khan's patient classification results before and after applying class data sampling techniques for all the used classifiers.

schemes of Khan *et al.*'s and Robinson's and the other computer-aided classification schemes based on the best six classification results were expanded using the JELEN_MERGE01 dataset after adjusting the class distribution.

	Average accuracy		Average		
Classification algorithms	**Previous**	**New**	**Sensitivity**	**Specificity**	**Precision**
SVM-OS	96.66%	**99.07%**	99.14%	**98.94%**	**98.95%**
DT+AdaBoost-OS	94.81%	98.67%	**99.74%**	97.59%	97.65%
RF-OS	95.72%	98.54%	99.49%	97.59%	97.64%
FFNN-OS	**97.58%**	97.61%	96.34%	98.89%	98.87%
DT-OS	92.40%	96.93%	97.59%	96.26%	96.32%
MLP-OS	93.45%	96.50%	94.93%	98.07%	98.02%

Table 20: Evaluation results (all the metrics calculated based on oversampled dataset except the previous value based on the imbalanced dataset) of the first best case classification for the six best classifiers using the JELEN_MERGE01 dataset for Khan's cytological grading frameworks. Best results indicated in bold. OS - Oversampled dataset. Previous: using imbalanced dataset, New: using balanced dataset.

Similarly, in Tables 22 and 23, for patient classification, the comparison between the best two schemes of Fisher's and Khan *et al.*'s and the other computer-aided classification schemes based on the best six classification results using the JELEN_MERGE01 dataset after adjusting class distribution is shown.

For the case and patient classification, as shown in the above Tables 20–23, the overall performance accuracies were improved for all six classifier results of the best two CA-CGSs. However, considering only the SVM classifier results with the oversampling technique for these original CA-CGSs indicates that the best two CA-CGSs with the rebalanced dataset were Khan *et al.*'s followed by Robinson's systems for the case classification and Robinson's followed by Khan *et al.*'s systems for the patient classification.

Since three CA-CGSs were modified by adding the low magnification features to each of Khan *et al.*'s, Fisher's and Taniguchi *et al.*'s systems as discussed in section 3.2, another evaluation among the nine

	Average accuracy		Average		
Classification algorithms	**Previous**	**New**	**Sensitivity**	**Specificity**	**Precision**
SVM-OS	**97.28**%	**98.94**%	98.85%	98.39%	98.42%
DT+AdaBoost-OS	96.22%	98.86%	**99.59**%	98.12%	98.14%
FFNN-OS	95.76%	98.62%	98.32%	98.92%	98.91%
RF-OS	95.97%	98.13%	99.19%	97.06%	97.13%
MLP-OS	95.76%	98.03%	96.94%	**99.12**%	**99.10**%
DT-OS	93.75%	97.04%	97.41%	96.66%	96.71%

Table 21: Evaluation results (all the metrics calculated based on oversampled dataset except the previous value based on the imbalanced dataset) of the second best case classification for the best six classifiers using the JELEN_MERGE01 dataset for Robinson's cytological grading frameworks. Best results indicated in bold. OS - Oversampled dataset. Previous: using imbalanced dataset, New: using balanced dataset.

	Average accuracy		Average		
Classification algorithms	**Previous**	**New**	**Sensitivity**	**Specificity**	**Precision**
DT+AdaBoost-OS	90.84%	**98.96**%	**99.87**%	**97.91**%	**98.21**%
SVM-OS	**92.69**%	96.05%	95.87%	96.25%	96.70%
RF-OS	91.24%	98.02%	98.24%	97.77%	98.06%
DT-OS	92.27%	96.18%	95.03%	97.50%	97.76%
FFNN-OS	85.92%	94.91%	94.23%	95.51%	94.88%
MLP-OS	76.71%	93.81%	94.86%	92.90%	92.18%

Table 22: Evaluation results (all the metrics calculated based on oversampled dataset except the previous value based on the imbalanced dataset) of the first best patient classification for the best six classifiers using the JELEN_MERGE01 dataset for Fisher's cytological grading frameworks. Best results indicated in bold. OS - Oversampled dataset. Previous: using imbalanced dataset, New: using balanced dataset.

	Average accuracy		Average		
Classification algorithms	**Previous**	**New**	**Sensitivity**	**Specificity**	**Precision**
DT+AdaBoost-OS	92.32%	**98.70**%	**99.39**%	97.91%	98.20%
RF-OS	93.74%	98.64%	98.78%	98.47%	98.67%
SVM-OS	96.24%	98.35%	98.66%	97.98%	98.24%
FFNN-OS	**97.11**%	97.50%	97.29%	97.69%	97.39%
MLP-OS	92.22%	96.79%	96.25%	97.27%	96.87%
DT-US	92.16%	96.11%	92.66%	**99.55**%	**99.58**%

Table 23: Evaluation results (all the metrics calculated based on oversampled dataset except the previous value based on the imbalanced dataset) of the second best patient classification results for the best six classifiers using the JELEN_MERGE01 dataset for Khan's cytological grading frameworks. Best results indicated in bold. US - Undersampled dataset. OS - Oversampled dataset. Previous: using imbalanced dataset, New: using balanced dataset.

CA-CGSs with the imbalanced data (see Tables 24 and 26) and the rebalanced JELEN_MERGE01 dataset was performed in this chapter. The evaluation was based only on the SVM classifier and the oversampling technique results, as together they achieved the best accuracy performances for almost all the CA-CGSs (see Tables 25 and 27). In table 25, for case classification, the CA-CGS based on the modified Khan *et al.*'s followed by the modified Taniguchi *et al.*'s systems achieved the two best accuracies among all nine CA-CGSs with the rebalanced JELEN_MERGE01 dataset. Similarly, in table 27, for patient classification, the CA-CGS based on modified Khan *et al.*'s followed by the modified Taniguchi *et al.*'s systems achieved the two best accuracies out of all nine CA-CGSs using the JELEN_MERGE01 dataset.

Cytological schemes	**Average accuracy**	**Average sensitivity**	**Average specificity**	**Average precision**
Robinson's *	97.28%	93.33%	98.27%	93.09%
Fisher's	93.35%	75.25%	97.84%	89.68%
Khan *et al.*'s	96.66%	88.68%	**98.64**%	**94.24**%
Taniguchi *et al.*'s	95.30%	84.94%	97.87%	90.92%
Mouriquand's	95.56%	88.68%	97.26%	89.09%
Howell's	85.64%	66.16%	90.47%	63.48%
Modified Fisher's	96.44%	88.89%	98.32%	92.93%
Modified Khan *et al.*'s ☆	**97.77**%	**95.65**%	98.29%	93.34%
Modified Taniguchi *et al.*'s	97.24%	93.03%	98.29%	93.16%

Table 24: Evaluation results of case classification using the SVM classifier on the imbalanced JELEN_MERGE01 dataset for all nine proposed CA-CGSs. Best results indicated in bold. . ☆ - The first best CA-CGS. * - The second best CA-CGS.

To evaluate the robustness of the modified three CA-CGSs as well as to determine the best CA-CGS among all nine CA-CGSs after adjusting the class distributions in the used dataset, the average of the four measurements, accuracy, sensitivity, specificity, and precision were estimated for 30 runs to get the optimized model result. Further, the confusion matrices, were calculated for the best case and patient classification systems among all nine proposed CA-CGSs as shown in Figure 52.

On the other hand, the final evaluation between the modified Khan *et al.*'s, modified Taniguchi *et*

Cytological schemes	Average accuracy	Average sensitivity	Average specificity	Average precision
The six original CA-CGSs with re-balanced dataset				
Robinson's	98.94%	**98.85%**	98.39%	98.42%
Fisher's	98.13%	98.29%	97.97%	97.98%
Khan *et al.*	99.07%	99.14%	98.94%	98.95%
Taniguchi *et al.*	98.06%	98.07%	99.14%	99.04%
Mouriquand's	98.16%	98.57%	99.14%	99.15%
Howell's	97.16%	99.07%	95.26%	95.44%
The three modified CA-CGSs with re-balanced dataset				
Modified Fisher's	99.17%	98.49%	99.85%	99.84%
Modified Khan *et al.*'s ☆	**99.83%**	99.69%	**99.97%**	**99.97%**
Modified Taniguchi *et al.* *	99.58%	99.82%	99.34%	99.34%

Table 25: Evaluation results of case classification based on the SVM-OS results using the rebalanced JELEN_MERGE01 dataset for all nine CA-CGSs. Best results indicated in bold. ☆ - The first best CA-CGS. * - The second best CA-CGS.

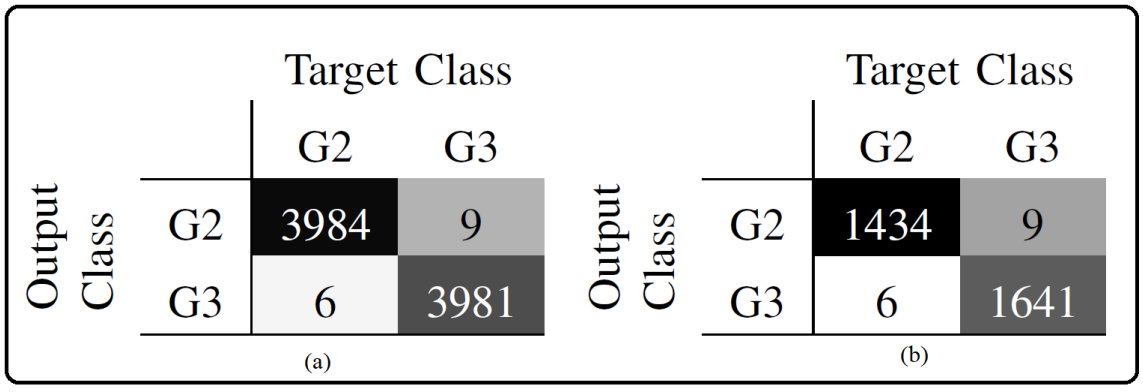


Figure 52: Example of the calculated confusion matrices for 30 runs for the best-modified system of the Khan *et al.*'s CA-CGS. (a) Case classification confusion matrix and (b) Patient classification confusion matrix.

Cytological schemes	**Average accuracy**	**Average sensitivity**	**Average specificity**	**Average precision**
Robinson's	95.29%	85.77%	98.26%	93.99%
Fisher's	92.69%	70.00%	99.79%	99.20%
Khan *et al.* *	96.24%	89.33%	98.40%	94.66%
Taniguchi *et al.*'s	94.23%	80.66%	98.47%	94.47%
Mouriquand's	95.02%	80.44%	99.58%	98.41%
Howell's	86.08%	61.33%	93.81%	75.91%
Modified Fisher's	95.71%	82.00%	**100**%	**100**%
Modified Khan *et al.*'s ☆	**96.50**%	**91.77**%	97.98%	93.50%
Modified Taniguchi *et al.*'s	95.29%	86.22%	98.12%	93.60%

Table 26: Evaluation results of patient classification using the SVM classifier on the imbalanced JELEN_MERGE01 dataset for all nine CA-CGSs. Best results indicated in bold. ☆ - The first best CA-CGS. * - The second best CA-CGS.

Cytological schemes	Average accuracy	Average sensitivity	Average specificity	Average precision
The six original CA-CGSs with re-balanced dataset				
Robinson's	98.86%	**99.63%**	97.98%	98.27%
Fisher's	96.05%	95.87%	96.25%	96.70%
Khan *et al.*	98.35%	98.66%	97.98%	98.24%
Taniguchi *et al.*'s	97.50%	96.66%	98.47%	98.64%
Mouriquand's	98.08%	98.09%	99.13%	99.14%
Howell's	97.12%	98.30%	95.76%	96.38%
The three modified CA-CGSs with re-balanced dataset				
Modified Fisher's	98.05%	96.36%	**100%**	**100%**
Modified Khan *et al.*'s ☆	**99.61%**	99.27%	**100%**	**100%**
Modified Taniguchi *et al.*'s *	98.99%	99.57%	98.33%	98.56%

Table 27: Evaluation results of patient classification based on the SVM-OS results using the rebalanced JELEN_MERGE01 dataset for all nine CA-CGSs. Best results indicated in bold. ☆ - The first best CA-CGS. * - The second best CA-CGS.

al.'s and modified Fisher's systems and their original versions was undergone in this chapter in terms of the used imbalanced and re-balanced JELEN_MERGE01 dataset (see Tables 28 and 29).

Imbalanced data cause biasing towards the majority class while the minority class is usually neglected, due to which the sensitivity and precision of the system is decreased (see Figures 53 and 54). To handle this issue, a rebalancing of the system was done using the oversampling method as it gave the best results out of the three selected methods: oversampling, undersampling, and RUSBoost. By doing so, for case classification, for all nine CA-CGSs, the overall accuracy was improved overall by 4.19% with a standard deviation of 3.66%. Howell's system's accuracy for case classification improved significantly by 13.47%. This means that this system was biased towards the majority class compared to the minority class. A similar trend was observed for patient classification in which the overall accuracy improved by 4.11% with a standard deviation of 2.70%, while Howell's accuracy significantly improved by 11.04%.

Cytological schemes	**Average accuracy**	**Average sensitivity**	**Average specificity**	**Average precision**
Original CA-CGSs with imbalanced dataset				
Fisher's	93.35%	75.25%	97.84%	89.68%
Khan *et al.*	96.66%	88.68%	**98.64**%	**94.24**%
Taniguchi *et al.*	95.30%	84.94%	97.87%	90.92%
Original CA-CGSs with rebalanced dataset				
Fisher's	98.13%	98.29%	97.97%	97.98%
Khan *et al.*	99.07%	99.14%	98.94%	98.95%
Taniguchi *et al.*	98.06%	98.07%	99.14%	99.04%
Modified CA-CGSs with imbalanced dataset				
Modified Fisher's	96.44%	88.89%	98.32%	92.93%
Modified Khan *et al.*'s	**97.77**%	**95.65**%	98.29%	93.34%
Modified Taniguchi *et al.*'s	97.24%	93.03%	98.29%	93.16%
Modified CA-CGSs with rebalanced dataset				
Modified Fisher's	99.17%	98.49%	99.85%	99.84%
Modified Khan *et al.*'s	**99.83**%	99.69%	**99.97**%	**99.97**%
Modified Taniguchi *et al.*'s	99.58%	99.82%	99.34%	99.34%

Table 28: Evaluation results of case classification based on the SVM for the three original and the three modified CA-CGSs before and after rebalancing the class distribution of used JELEN_MERGE01 dataset. Best results indicated in bold.

Cytological schemes	**Average accuracy**	**Average sensitivity**	**Average specificity**	**Average precision**
Original CA-CGSs with imbalanced dataset				
Fisher's	92.69%	70.00%	**99.79**%	**99.20**%
Khan *et al.*	**96.24**%	**89.33**%	98.40%	94.66%
Taniguchi *et al.*	94.23%	80.66%	98.47%	94.47%
Original CA-CGSs with rebalanced dataset				
Fisher's	96.05%	95.87%	96.25%	96.70%
Khan *et al.*	98.35%	98.66%	97.98%	98.24%
Taniguchi *et al.*	97.50%	96.66%	98.47%	98.64%
Modified CA-CGSs with imbalanced dataset				
Modified Fisher's	95.71%	82.00%	**100**%	**100**%
Modified Khan *et al.*'s	**96.50**%	**91.77**%	97.98%	93.50%
Modified Taniguchi *et al.*'s	95.29%	86.22%	98.12%	93.60%
Modified CA-CGSs with rebalanced dataset				
Modified Fisher's	98.05%	96.36%	**100**%	**100**%
Modified Khan *et al.*'s	**99.61**%	99.27%	**100**%	**100**%
Modified Taniguchi *et al.*'s	98.99%	99.57%	98.33%	98.56%

Table 29: Evaluation results of patient classification based on the SVM for the three original and the three modified CA-CGSs before and after rebalancing the class distribution of used JELEN_MERGE01 dataset. Best results indicated in bold.

Another major concern is the sensitivity of the system which improved by 13.13% overall with a standard deviation of 9.21% for all nine CA-CGSs. For case classification, two interesting systems in which the sensitivity significantly improved are Fisher's and Howell's. Fisher's and Howell's improved as they are biased towards the majority class, so by rebalancing the data, they lost the bias and consequently, the sensitivity rate improved. The same phenomenon was observed for patient classification in which Fisher's and Howell's showed a significant improvement of the overall sensitivity of the CA-CGSs by 17.32% with a standard deviation of 9.07% (see Figures 55 and 56). There was no significant improvement in specificity of the case and patient classifications for CA-CGSs which is expected, because the specificity is the measure of the majority class for which there is enough data for the training. Nonetheless, there is a slight decrease of the specificity in Robinson's, Fisher's, and Khan's systems for patient classification because of the discriminative abilities of some of the estimated features or due to the data overlapping characteristics problem (samples from different classes share the same characteristics) for these systems. For precision measurement, there is an overall improvement of 9.81% and 5.68% with a standard deviation of 8.46% and 6.30% for the case and patient classification, respectively. In this case, also, Howell's system showed a major improvement by 31.96% and 20.47%, respectively. However, for patient classification, the specificity in the case of Fisher's system decreased by 2.5%. This could again be due to data overlapping characteristics and the type of features related to the cytological malignancy criteria of the system.

Cytological schemes	Average accuracy	Sensitivity	Specificity	Precision
Robinson's *	97.28%	93.33%	98.27%	93.09%
Fisher's	93.35%	75.25%	97.84%	89.68%
Khan *et al.*'s	96.66%	88.68%	**98.64%**	**94.24%**
Taniguchi *et al.*'s	95.30%	84.94%	97.87%	90.92%
Mouriquand's	95.56%	88.68%	97.26%	89.09%
Howell's	85.64%	66.16%	90.47%	63.48%
Modified Fisher's	96.44%	88.89%	98.32%	92.93%
Modified Khan *et al.*'s ☆	**97.77%**	**95.65%**	98.29%	93.34%
Modified Taniguchi *et al.*'s	97.24%	93.03%	98.29%	93.16%

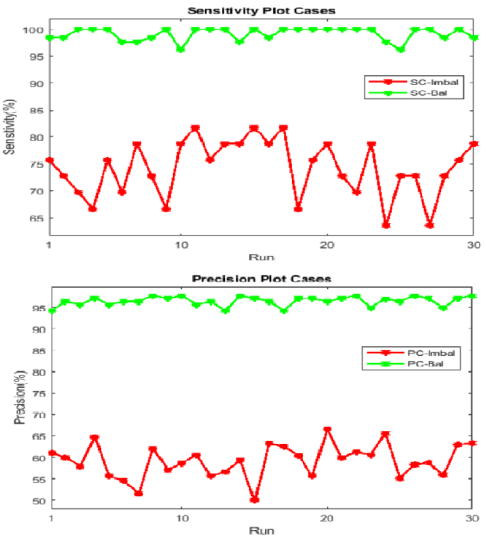


Figure 53: Example of the sensitivity and precision rates of the Fisher's and Howell's systems for the case classification using the imbalanced JELEN_MERGE01 dataset. SC-Imbal: The imbalanced dataset sensitivity rates for the case classification. SC-Bal: The balanced dataset sensitivity rates for the case classification.

Cytological schemes	Average accuracy	Sensitivity	Specificity	Precision
The six original CGSs with re-balanced dataset				
Robinson's	98.86%	**99.63%**	97.98%	98.27%
Fisher's	96.05%	95.87%	96.25%	96.70%
Khan *et al.*	98.35%	98.66%	97.98%	98.24%
Taniguchi *et al.*'s	97.50%	96.66%	98.47%	98.64%
Mouriquand's	98.08%	98.09%	99.13%	99.14%
Howell's	97.12%	98.30%	95.76%	96.38%
The three modified CGSs with re-balanced dataset				
Modified Fisher's	98.05%	96.36%	**100%**	**100%**
Modified Khan *et al.*'s ☆	**99.61%**	99.27%	**100%**	**100%**
Modified Taniguchi *et al.*'s *	98.99%	99.57%	98.33%	98.56%

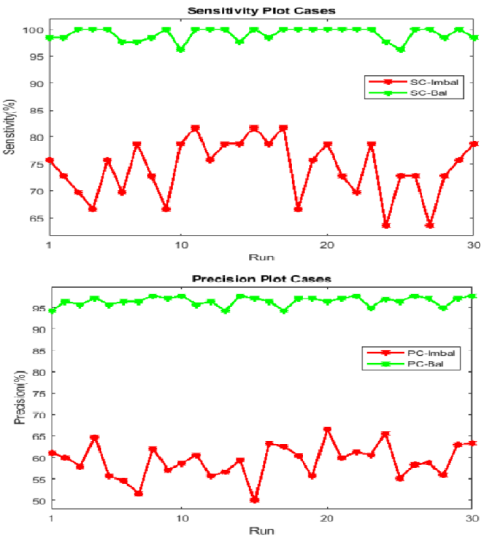


Figure 54: Example of the sensitivity and precision rates of the Fisher's and Howell's systems for the case classification using the balanced JELEN_MERGE01 dataset. SC-Imbal: The imbalanced dataset sensitivity rates for the case classification. SC-Bal: The balanced dataset sensitivity rates for the case classification.

Cytological schemes	Average accuracy	Sensitivity	Specificity	Precision
Robinson's	95.29%	85.77%	98.26%	93.99%
Fisher's	92.69%	70.00%	99.79%	99.20%
Khan *et al.* *	96.24%	89.33%	98.40%	94.66%
Taniguchi *et al.*'s	94.23%	80.66%	98.47%	94.47%
Mouriquand's	95.02%	80.44%	99.58%	98.41%
Howell's	86.08%	61.33%	93.81%	75.91%
Modified Fisher's	95.71%	82.00%	**100%**	**100%**
Modified Khan *et al.*'s ☆	**96.50%**	**91.77%**	97.98%	93.50%
Modified Taniguchi *et al.*'s	95.29%	86.22%	98.12%	93.60%

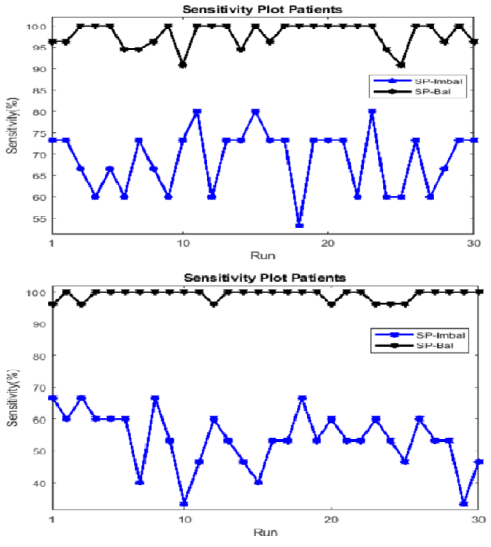


Figure 55: Example of the sensitivity rates of the Fisher's and Howell's systems for the patient classification using the imbalanced JELEN_MERGE01 dataset. SC-Imbal: The imbalanced dataset sensitivity rates for the case classification. SC-Bal: The balanced dataset sensitivity rates for the case classification.

Cytological schemes	Average accuracy	Sensitivity	Specificity	Precision
The six original CGSs with re-balanced dataset				
Robinson's	98.86%	**99.63%**	97.98%	98.27%
Fisher's	96.05%	95.87%	96.25%	96.70%
Khan *et al.*	98.35%	98.66%	97.98%	98.24%
Taniguchi *et al.*'s	97.50%	96.66%	98.47%	98.64%
Mouriquand's	98.08%	98.09%	99.13%	99.14%
Howell's	97.12%	98.30%	95.76%	96.38%
The three modified CGSs with re-balanced dataset				
Modified Fisher's	98.05%	96.36%	**100%**	**100%**
Modified Khan *et al.*'s ☆	**99.61%**	99.27%	**100%**	**100%**
Modified Taniguchi *et al.*'s *	98.99%	99.57%	98.33%	98.56%

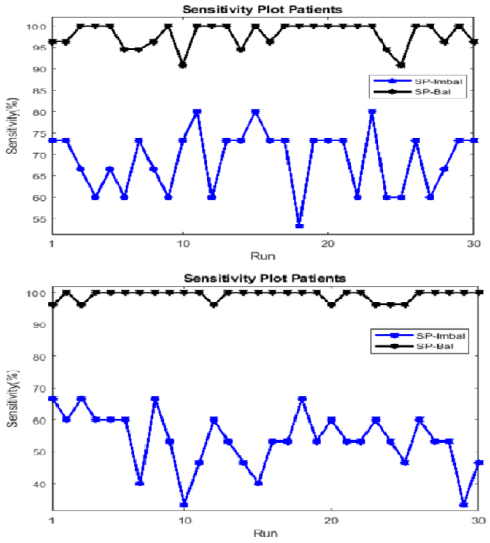


Figure 56: Example of the sensitivity rates of the Fisher's and Howell's systems for the patient classification using the balanced JELEN_MERGE01 dataset. SC-Imbal: The imbalanced dataset sensitivity rates for the case classification. SC-Bal: The balanced dataset sensitivity rates for the case classification.

SC-Bal: The balanced dataset sensitivity rates for the case classification.

4.3 Conclusions

In this chapter, we considered the usage of the RUSBoost ensemble-learning algorithm and two data sampling techniques to overcome the difficulty that occurs with the imbalanced classification problem of breast cancer cytological malignancy. Following a detailed comparison between the two data sampling techniques and the RUSBoost approach, the experimental results showed that the data sampling techniques outperformed the average accuracy of the RUSBoost approach by at least 8% for all the original six cytological grading systems. Furthermore, it was found that data sampling techniques, particularly oversampling, were able to enhance the overall performance accuracy of all nine proposed computer-aided cytological grading schemes. More precisely, the performance accuracy of all nine CA-CGSs was improved by 4.19% overall with a standard deviation of 3.66% for case classification and 4.11% with a standard deviation of 2.70% for patient classification based on the SVM classifier.

Considering only the six original CA-CGSs, Khan *et al.*'s and Robinson's systems achieved the best performance accuracies for case classification, while Fisher's and Khan *et al.*'s achieved the best performance accuracies for patient classification. On the other hand, considering all nine CA-CGSs, the best two accuracies for case classification ware 99.83% and 99.58%, obtained by the

SVM classifier with the oversampling technique for computer-aided versions of the modified Khan *et al.*'s and modified Taniguchi *et al.*'s, respectively. For patient classification, the obtained accuracies were 99.61% and 98.99%, also obtained by the SVM classifier with the oversampling technique for the computer-aided version of the modified Khan *et al.*'s and modified Taniguchi *et al.* systems, respectively.

Moreover, the sensitivity rate of Fisher's and Howell's systems witnessed a significant improvement after rebalancing the class distribution, while, for patient classification, there a slight decrease of specificity in Robinson's, Fisher's, and Khan's systems which might be due to the features correlation or the discriminative abilities of some of the estimated features for these systems. Furthermore, the precision measurement improved for the case and patient classification for almost all the proposed CA-CGSs.

Chapter 5

Computer-aided cytological malignancy grading systems for fine needle aspiration biopsies of breast cancer based on convolutional neural networks

Recently, DL has emerged as a promising technique, surpassing common approaches, solving challenging problems in several areas such as image classification, speech recognition, and object detection [93, 94, 95, 96]. DL produces information based on the automatic extraction of the elementary characteristics such as color, texture, and motion of an object in an image which is not user-dependent. Therefore, the performance of these systems does not degrade for different users since feature extraction is automatic and not subjective. On the other hand, traditional ML algorithms heavily rely on the handcrafted estimated features that require domain expertise. Although the feature extraction task is automatic, not manual, its robustness is still subjective to different users with respect to their domain knowledge. More precisely, with respect to the underlying problem for this study, to determine the malignancy grade of FNA images, in each of the proposed CA-CGS, an important step is to accurately segment cell nuclei regions for feature extraction and subsequent classification purposes (as we discussed in chapter 3). Unfortunately, traditional analysis

for cytological images of FNA biopsies of tissue is challenging due to the fact that cells tend to naturally generate clusters of connected nuclei regions, or nests. Thus, it is difficult to achieve accurate separated nuclei regions as required for accurate feature calculation. Furthermore, the malignancy grading tasks as studied in these papers [83, 125] are defined according to specific histopathology/-cytopathology grading systems that require careful analysis of FNA images to determine specific nuclear and cellular characteristics that reflect the biological behavior of cancerous cells. Thus, this challenge of determining the most appropriate handcrafted features to be used by traditional classifiers has further complicated the task of automating the diagnosis and grading systems to assist the specialists in their manual examinations of FNA biopsies.

Meanwhile, DL focuses on learning multiple features of images which aid in tasks such as image classification and face recognition. These features are abstracted from different levels of the image representation coming from the DL network. For instance, an image can be shown as a vector of intensity values for pixels or regions of interest. A key significance of DL methods is automating manual work such as handcrafted feature engineering using efficient networks [126]. A CNN model is basically a deep neural network which consists of multiple hidden layers of convolution and pooling functions in addition to the activation function. It does not require an image segmentation process and efficiently replaces the handcrafted features with efficient feature learning layers and hierarchical feature extraction techniques [126]. By making use of CNN models, we can automatically extract features and abstractions from the underlying data without the need of consuming a lot of time and effort to accomplish different fundamental tasks of the image analysis that are usually used by traditional classification models. Despite the significant performance that have been obtained using CNN models for the problem of computer-aided diagnosis and malignancy grading systems for breast cancer using various modalities such as histopathology [127, 128, 129, 130], ultrasonography [131], MRI [132], and mammography [133], the malignancy grading problem for cytological images of breast cancer, on the other hand, has been much less studied using CNN models. To the best of our knowledge, the existing literature is more limited for this problem and has been done primarily using conventional approaches of image processing and traditional machine learning algorithms.

In this chapter, we address the gap in existing literature by proposing the first CNN-CGS for FNA biopsies of breast cancer. To achieve our goal, a variety of pre-trained CNN architectures such as AlexNet, GoogleNet-inception-v3, and ResNet were used in this thesis. Pre-trained CNN models were used, as training CNN from scratch requires a massive amount of data samples, which are required for optimal performance of a CNN model. However, for datasets with few data samples, as is the case with the datasets used in this study, pre-trained CNNs [134] with weights previously trained using the ImageNet dataset [135] can be then used to initialize the lower layers of our CNN

models and the model can then be fine-tuned on the target dataset. The main concern of this chapter is to investigate the discriminatory power of the CNN-based features for the cytological grading problem for breast cancer and compare it with the discriminatory power of the previously extracted handcrafted-based features and to determine the best performance model of CNN architectures for the underlying the cytological grading problem for FNA biopsies. Here, different evaluation methods were used to evaluate the robustness of the proposed CNN-based models.

5.1 The methodology of the proposed frameworks

With the aim of studying the impact of CNN-based extracted features versus handcrafted extracted features on the overall performance accuracy of the malignancy grading problem for FNA biopsies for breast cancer, a variety of CNN architectures were studied to propose computer-aided cytological malignancy grading systems (CNN-CMGs) for FNA biopsies of breast cancer. Formerly, six CA-CGSs were proposed to determine the malignancy grades for FNA biopsies of breast cancer based on estimating different morphologic, polymorphic and textural handcrafted extracted features. Particularly, each system was tailored to follow the cytological characteristics and a scoring system was defined for each of the considered six grading schemes. According to the obtained results, the best results using imbalanced dataset were obtained for the SVM classifier for computer-aided versions of the Robinson's and Khan *et al.*'s cytological grading systems with accuracies of 97.28% and 96.66% for case classification and 95.29% and 96.24% for patient classification, respectively (see tables 30 and 31). On the other hand, after adjusting the class distribution among the proposed six CGSs, for the case and patient classification, the best results using the rebalanced dataset were obtained by the SVM classifier and the oversampled technique for the computer-aided version of Khan *et al.*'s and Fisher's CGSs with accuracies of 99.07% and 98.13%, respectively.

In this thesis, due to the small size of the collected training dataset, a transfer learning and fine-tuning method were performed on different types of pre-trained image classification networks (AlexNet, Inception-v3, and ResNet) to propose CNN-CMGs for FNA biopsies of breast cancer. The used networks have been previously trained to extract powerful and meaningful features from different types of natural images of ImageNet. Thus, they were used as a starting point to learn the underlying classification task for this chapter (malignancy grading for FNA images).

Further, we considered four types of inputs of images, all taken from the JELEN_MERGE dataset, for the training and testing of the CNN models used:

1. *High magnification (400x) images only.* High magnification images capture the cellular and

Cytological schemes	Average accuracy	Average sensitivity	Average specificity	Average precision
Robinson's	**97.28**%	**93.33**%	98.27%	93.09%
Khan *et al.*	96.66%	88.68%	**98.64**%	**94.24**%
Mouriquand's	95.56%	88.68%	97.26%	89.09%
Taniguchi *et al.*	95.30%	84.94%	97.87%	90.92%
Fisher's	93.35%	75.25%	97.84%	89.68%
Howell's	83.87%	59.89%	89.82%	59.62%

Table 30: Evaluation results on the case classification based on the SVM classifier on the imbalanced JELEN_MERGE01 dataset for all six original CA-CGSs from section 2 of chapter 3. Best results indicated in bold.

Cytological schemes	Average accuracy	Average sensitivity	Average specificity	Average precision
Khan *et al.*	**96.24**%	**89.33**%	98.40%	94.66%
Robinson's	95.29%	85.77%	98.26%	93.99%
Mouriquand's	95.02%	80.44%	99.58%	98.41%
Taniguchi *et al.*'s	94.23%	80.66%	98.47%	94.47%
Fisher's	92.69%	70.00%	**99.79**%	**99.20**%
Howell's	84.60%	54.88%	93.88%	74.59%

Table 31: Evaluation results on the patient classification based on the SVM classifier on the imbalanced JELEN_MERGE01 dataset for all six original CA-CGSs from section 2 of chapter 3. Best results indicated in bold.

nuclear characteristics in the FNA cytological images. These characteristics are strong malignancy indicators since cancerous cells tend to be irregularly shaped with greater size variance than normal cells. In particular, three of the six cytological grading schemes listed in Tables 30 and 31 use only high magnification images to assign malignancy grades.

2. *Low magnification (100x) images only.* Low magnification images capture larger scale characteristics, such as the dispersion level of cells in FNA images which represents a good malignancy indicator since cancerous cells tend to be more widely dispersed with less cellular structure.

3. *Combined low and high magnification images.* Here, FNA images of both magnifications (100x and 400x) were combined while ignoring the type of magnification for each image. A similar approach was used by Bayramoglu *et al.* [136] where they proposed a magnification independent approach for malignancy diagnosis of the histopathological breast cancer images of the BreaKHis data set [137], ignoring the magnification information of the images and training a unique CNN classifier for all magnifications (40x, 100x, 200x, and 400x).

4. *Concatenated images of pairs of 100x and 400x images.* Here, pairs of 100x and 400x magnification images, each pair of images coming from the same patient, are concatenated and scaled to the same size as the original images (see Figure 57 for example and Tables 32 and 33 for images sizes used). The merging of the two magnifications is similar to the case classification used in our previous work [125] and related work on histological images [21].

In all the above types of inputs, the patient label information was retained.

However, for unbalanced datasets such as our datasets, it is useful to investigate properly re-balancing the distribution of the classes. Therefore, to handle this crucial problem, the class distribution in our dataset was adjusted to improve the performance of the proposed CNN-CMGs. For this purpose, three approaches were examined, in addition to using the original unbalanced datasets. In the first approach, the data augmentation process was applied to transform the training data for each epoch. This approach helps prevent the CNN from overfitting as well as help the model to learn the exact details of the training images. To apply augmentation, the training images were randomly flipped horizontally (left-right direction) with 50% probability and randomly translated up to 30 pixels horizontally and vertically. Also, the images were randomly scaled up to 10% horizontally and vertically. In the second and third approaches, two data sampling techniques were utilized, oversampling (minority class) and undersampling (majority class), to achieve a balanced (50:50) data distribution. For the oversampling technique, to overcome imbalanced class data distribution, randomly duplicated samples were added to the minority class (G3 in our datasets). The undersampling technique, on the other hand, handles this problem by randomly eliminating samples from the

majority class (G2 in our datasets).

5.2 Classification Tasks

The classification task of the presented CNN-CMGs for FNA biopsies of breast cancer based on different pre-trained CNN architectures is discussed in this section. To determine the malignancy level of FNA biopsies, pre-trained CNN models were used to perform the classification directly from the images. The CNN models take an image as an input and, as an output, respond with a label associated with the objects in the image as well as the probabilities for each of the object categories based on implicitly extracted features by the networks. Since pre-trained CNN models were used in this study, the feature extraction task becomes easier and faster to accomplish compared to the time and effort required to train a full network from scratch. As previously mentioned, three different CNN models were used in this thesis. First, we used AlexNet (2012) [69], a network consisting of 8 layers in-depth, 5 of which are convolutional layers, which significantly outperformed all the prior competitors and won the ILSVRC-2012 competition by achieving a top-5 test error rate of 15.3%. Then we used an Inception-v3 model of GoogleNet, featuring 48 layers in depth. GoogleNet (2014) [70], which consists of 22 layers in-depth, won the ILSVRC 2014 competition by achieving a top-5 error rate of 6.67%. Finally, we used the Residual Neural Network (ResNet) (2015) [75], which achieved a top-5 error rate of 3.57% and won the 1st place on the ILSVRC-2015 classification task. Specifically, for this thesis, three different versions of ResNet were used which were 18, 50 and 101 layers in-depth.

According to pathology-based studies, some of the published cytological schemes have only been used on high magnifications images of 400x power to determine the malignancy grade (low, intermediate or high) of breast cancer. However, other cytological grading schemes have been provided a sequence of images with different magnifications to determine the malignancy grade of breast cancer. It was not required to mimic any of the published pathology-based cytological grading schemes to develop these particular CNN-based systems; however, the importance of the image magnifications on the malignancy grading process in the current grading systems, as is the case with the handcrafted-based proposed CGSs in chapter 3, was considered. Thus, the proposed CNN-CMGs consists of two classification schemes: Image classification and patient classification [103, 52]. Image classification involves a single feature calculation, while patient classification involves multiple feature calculations across multiple images with two different magnifications that belong to a certain patient. Precisely, the image classification task included three different methods of classification to determine which method would give the best accuracy performance for the proposed CNN-CMG systems. In the first method, both magnifications of FNA images (100x and 400x) were used to classify each case,

but the relationship between each pair of images present in our dataset was not considered in this method. Thus, each image, whether 100x or 400x, was treated as a separated case. In the second method, since some of the cytological grading schemes use only high magnification images to assign malignancy grades for a certain case, only the high magnification images were used to determine the malignancy level for each case belonging to a certain patient. In the third method, due to the importance of the low magnification features in determining the dispersion level of cancerous cells in FNA images (which represent a good malignancy indicator in most CGSs), only the low magnification images were used to determine the malignancy level for each case that belong to a certain patient. On the other hand, in the patient classification, to classify a certain patient as G2 or G3, the final classification results are achieved by majority voting of the image classification results (combined low and high, low only or high only) for a certain patient. Thus, if at least 50% of the images that belong to a certain patient are classified as G2, the final classification result of that patient is considered as G2; otherwise, it is classified as G3.

In this chapter, due to the fact that each pair of images with 100x and 400x magnifications creates a single case of multiple cases belonging to a certain patient in our used dataset, primitive attempts were performed to classify each case (pair of images) that belongs to a certain patient individually. To do so, each pair of images with 100x and 400x magnification were concatenated to create a single case. Then, the resulting concatenated images out of each pair were resized to the input image layer of each of the used CNN models (see Figure 57). Next, each concatenated pair of images with 100x and 400x (representing a single case) were classified separately. The patient classification in this scenario is obtained by taking the majority voting of the case classification results that belong to each patient instead of taking the majority voting of the image classification results that belong to each patient, as the scenario in the previous two methods.

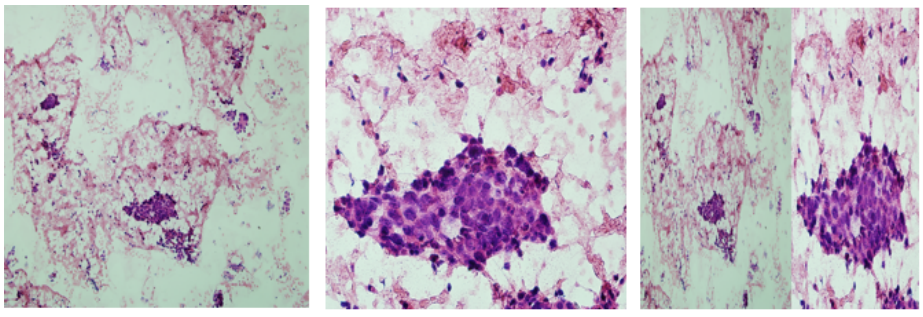


Figure 57: Example images of one case, (a) Low magnification (100x), (b) High magnification (400x) and (c) concatenated pair of images, from JELEN18 dataset.

In terms of classification accuracy evaluation for the used pre-trained networks, the standard top-1

accuracy (to estimate the correct predicted samples to the total number of the predicted samples) has been used as the standard evaluation method for all the examined pre-trained CNN models in this chapter. However, if the dataset is unbalanced like our dataset, it is better to properly re-balance the distribution of the classes as well as include other evaluation methods such as sensitivity, specificity, etc. Therefore, to handle this crucial problem, the class distribution in our dataset was adjusted to improve the performance of the proposed CNN-CMGs. To do that, three approaches were examined for this purpose. In the first approach, the data augmentation process was applied to transform the training data for each epoch. This approach protects the network from overfitting problems and it helps the model memorize the exact details of the training images. The training images were randomly flipped along the vertical axis and randomly translated up to 30 pixels horizontally and vertically. Meanwhile, in the second and third approaches, two data sampling techniques, oversampling (minority class) and undersampling (majority class), were utilized to achieve a balanced (50:50) data distribution and to therefore enhance the performance of the proposed grading systems. The oversampling technique overcomes the imbalanced class data distribution by adding samples to the minority class (G3 in our dataset) by either duplicating the samples or adding new samples, whereas the undersampling technique handles this problem by eliminating samples from the majority class (G2 in our dataset).

5.3 Experimental Results

In this section, simulation results for the proposed computer-aided malignancy grading framework for FNA biopsies of breast cancer are discussed. The presented malignancy grading system is based on five different pre-trained CNN models that were retrained using two magnifications of FNA biopsy images. To retrain these convolution networks, a set of training options were determined using stochastic gradient descent with momentum. This algorithm is used to update the weight and bias parameters of the networks during the learning process to minimize the loss function by gradually moving in the direction of the negative gradient of the loss. Further, to get rid of local minima of the errors, a momentum term was added to find the global minima such that the accuracy of the networks is improved. The selected training options were determined carefully according to the training process results for each of the examined CNN model as well as the percentage of data division (training, testing, and validation subsets). In Table 32, the tuning options were given for each CNN model with 70% training set and 30% validation set of data division, while in Table 33, the tuning options were given for each CNN model with 50% training set and 50% validation set of data division. However, the achieved results from the five CNN models show that 70% training set and 30% validation set of data division gave the best classification results, thus, this percentage of

data division were considered and their results were reported in this chapter. Figure 58 composed of the final and full training options that were used for the examined pre-trained CNN models on 70% training set and 30% validation set of data division.

CNN models	**Input image size**	**Mini-batch size**	**Max-epoch size**	**Initial leraning rate**
Inception-v3	224x224x3	10	20	1e-3
ResNet101	224x224x3	10	6	1e-4
ResNet50	224x224x3	10	6	1e-4
ResNet18	224x224x3	10	6	1e-4
AlexNet	227x227x3	100	20	1e-3

Table 32: Training options for the used pre-trained CNN models using 70% training and 30% testing sets of JELEN_MERGE02 dataset.

CNN models	**Input image size**	**Mini-batch size**	**Max-epoch size**	**Initial leraning rate**
Inception-v3	224x224x3	10	6	1e-4
ResNet101	224x224x3	10	15	1e-6
ResNet50	224x224x3	10	15	1e-6
ResNet18	224x224x3	10	15	1e-6
AlexNet	227x227x3	100	20	1e-3

Table 33: Training options for the used pre-trained CNN models using 50% training and 50% testing sets of JELEN_MERGE02 dataset.

Since we used pre-trained networks with the transfer learning process in this thesis, we fine-tuned these networks to learn our new classification task using FNA images. This method is much faster and easier to accomplish than constructing and training a convolutional network from scratch, especially for small datasets such as in our case. To carry out the transfer learning, for each convolutional network model used, the early layers that already learned low-level features such as edges, blobs, and colors were maintained. The final layers that learned task-specific features (high-level features) were replaced with new layers in order to learn new features related to the problem at hand (malignancy

CNN models	Training Values for each CNN									
	Input image size	Mini-batch size	Max-epoch size	Momentum	Weight decay	Validation frequency	Learning Rate			W-B-factor
							Initial	Drop factor	Number Epochs	
Inception-v3	224x224x3	10	6	0.9	1e-4	3	3e-4	n/a	n/a	10
ResNet101	224x224x3	10	6	0.9	1e-4	3	1e-4	n/a	n/a	10
ResNet50	224x224x3	10	6	0.9	1e-4	3	1e-4	n/a	n/a	10
ResNet18	224x224x3	10	6	0.9	1e-4	3	1e-4	n/a	n/a	10
AlexNet	227x227x3	100	20	0.9	0.004	3	1e-3	0.1	8	0

Figure 58: The final and full training options for the used pre-trained CNN models using the JELEN_MERGE02 dataset. The learning Rate values indicate initial learning rates. The default values of the momentum and weight decay for the stochastic gradient descent with momentum were used with all the examined CNN models since they gave the best performance. For Alexnet, we set weight decay 0.004, initial learning rate 1e-3, and drop factor 0.1 to decrease the learning rate by it every 8 epochs (epoch consists of 3 iterations) during the training. The ***n/a*** values indicate that the learning rate remains ***constant*** throughout the training. For these models, to learn faster in the newly added layers as compared to the transferred layers, we multiplied the W-B-factor 10 by the global learning rate to determine the learning rate for the weights and biases for the new fully connected layers.

grading) using our FNA images. Then, the number of classes was updated according to our new dataset used in this thesis. Next, the training options for each pre-trained network were determined according to the information details in Tables 32 and 33. Further, some earlier layers were frozen for some of the used CNN models by setting the learning rates in those layers to zero to speed up the training process. Freezing is a beneficial procedure for pre-trained networks because it saves a lot of training time and frozen layers won't change any values during network training (their weights are calculated and trained previously) (see Figure 59). To test the method proposed in this chapter, we used the k-fold cross-validation technique, where k was set to $k = 5$, to divide the raw dataset into training and test sets. Once again, we didn't take patient information into account when dividing cases into training and testing due to the limitations associated with the datasets (see Experimental Results in Chapter 3). Further, the training set was manually divided into 70% training and 30% validation or 50% training and 50% validation subsets with the aim of examining which data division yields the best performance. Moreover, due to the fact that by using the k-fold cross-validation technique the samples have been divided randomly into training and testing subsets with each fold, 10 runs for each used CNN model were performed to optimize the obtained results. Additionally, the 95% confidence interval was calculated, using the Student's t-distribution, for the obtained classification accuracies, to provide a robust indication of the performance abilities of these CNN-based malignancy grading frameworks.

The proposed CNN-CMGs were used for two classification schemes: image classification and patient classification [103, 52]. Image classification, performed during each fold of an experimental run,

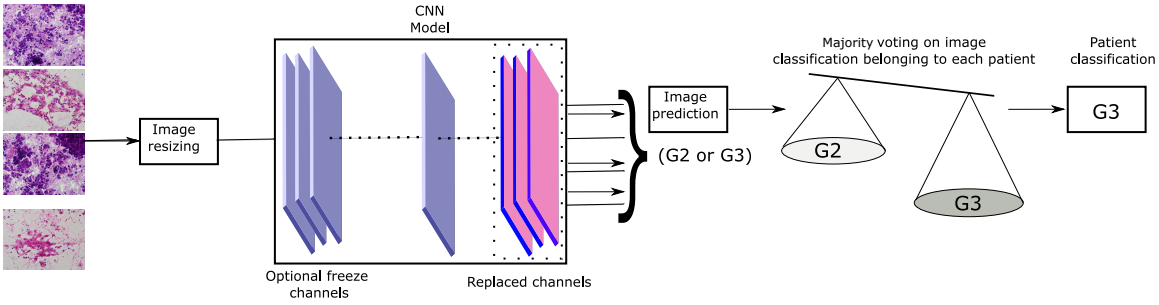


Figure 59: Overview of the workflow for the cytological malignancy grading systems (CNN-CMGs) based on CNN models.

involved the CNN-CMG systems assigning the input image output of either G2 or G3. On the other hand, for the patient classification to classify a certain patient as G2 or G3, the final classification results were achieved by majority voting of the image classification results for all the input images belonging to that patient. The patient classification was performed at the end of each experimental run. Thus, if at least 50% of the images that belong to the patient were classified as G2, the final classification result of that patient is considered as G2; otherwise, they were classified as G3.

In terms of classification accuracy evaluation for the used networks, the standard top-1 accuracy (estimated by the correct predicted samples to the total number of the predicted samples) has been used as the standard evaluation method for all the examined CNN-CMGs models in this study. In addition, we computed three measures from binary classification where we considered the G2 classification as the positive class and the G3 classification as the negative class. These measures are i) sensitivity—the proportion of the correct G2 classifications over all G2 classifications; specificity—the proportion of the correct G3 classifications over all G3 classifications; and precision—the proportion of the correct G2 classifications over all correct G2 and G3 classifications. To evaluate the robustness of the CNN-CMGS models, we computed the overall confusion matrix for each run of five folds while computing the two classification schemes (image and patient).

To evaluate the robustness of the proposed frameworks, the overall confusion matrices for each run per five folds were estimated to compare the two classification schemes (image and patient) as well as to investigate the misclassified images and patients for each run with respect to the malignancy grading task (see Figure 61). Also, the average of the accuracy per 10 runs, as well as the three other measures, that are, the sensitivity, specificity, and precision rates for each of the utilized pre-trained CNN models in the chapter, were computed.

In figure 61-(a), the diagonal cells to the right (451 and 74) in the matrix include the number of correct (TN and TP) classification samples by the GoogleNet-inception-v3 network, whereas

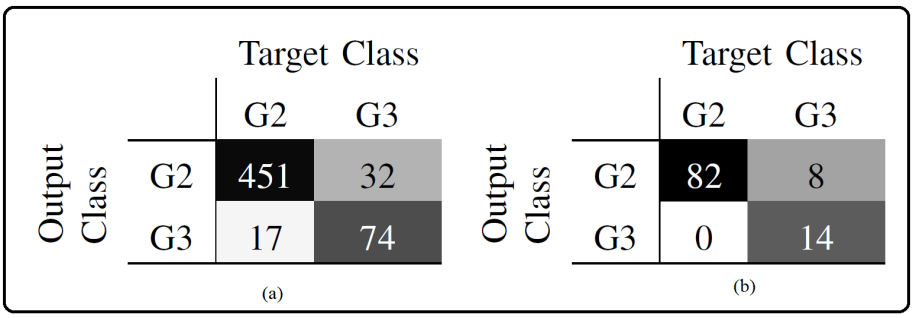


Figure 60: Example of the calculated confusion matrices for one run per 5 fold cross-validation. (a) Image classification confusion matrix and (b) Patient classification confusion matrix for CNN-CMG system based on GoogleNet-inception-v3 with imbalanced JELEN_MERGE02 dataset .

the diagonal cells to the left (32 and 17) in the matrix include the number of incorrect (FP and FN) classification samples by the GoogleNet-inception-v3 network. For instance, in this image confusion matrix, of all the classified FNA images, 451 (78.6%) images out of all 574 images were correctly classified as malignancy grade G2. Similarly, 74 (12.9%) samples were correctly classified as malignancy grade G3 of all used images. Also, 32 (5.6%) of the malignancy grade G3 images were incorrectly classified as G2 and 17 (12.9%) of the G2 images were incorrectly classified as G3. The same illustration applies to the patient confusion matrix results.

In terms of CNN results per computer-aided cytological grading system (CGS), four comparisons were performed while taking into account the problem of imbalanced class distribution in our used dataset. The first comparison was performed to measure the performance ability of the proposed CNN-CMGSs using the original imbalanced JELEN_MERGE02 dataset for both image and patient classification tasks. The three other comparisons were performed to measure the impact of the data augmentation and data sampling (undersampling and oversampling) approaches on improving the performance ability of the proposed CMGSs, designed based on pre-trained CNN models.

For the first image classification pattern (combined 100x and 400x), the results obtained from the pre-trained CNN using the combined low and high magnification images of the original imbalanced JELEN_MERGE02 dataset illustrate that the GoogleNet-inception-v3 networks performed better than the other CNN models. For image classification, in Table 34, beyond the best average accuracy and sensitivity rates, GoogleNet-inception-v3 also almost shared the best precision rate with ResNet101, while for patient classification, in Table 35, GoogleNet-inception-v3 achieved the best accuracy, sensitivity, and precision rates once again, sharing the best specificity with ResNet101, at 55% which is quite a low sensitivity due to the imbalanced dataset.

CNN models	Average accuracy	Average sensitivity	Average specificity	Average precision
Inception-v3	**89.98**%	**95.98**%	63.49%	92.08%
ResNet101	87.78%	92.60%	**66.50**%	**92.45**%
ResNet50	87.03%	92.79%	61.60%	91.45%
ResNet18	86.89%	94.40%	53.77%	90.06%
AlexNet	85.35%	93.03%	53.12%	89.94%

Table 34: Evaluation results on image classification of the proposed computer-aided CNN-CMGSs using the five examined CNN networks on the combined low and high magnification images of the original imbalanced JELEN_MERGE02 dataset. Best results indicated in bold.

CNN models	Average accuracy	Average sensitivity	Average specificity	Average precision
Inception-v3	**90.48**%	**100**%	**55.00**%	**89.29**%
ResNet101	88.17%	97.07%	**55.00**%	88.97%
ResNet50	86.73%	97.56%	46.36%	87.17%
ResNet18	86.92%	99.14%	41.36%	89.38%
AlexNet	86.35%	98.78%	39.54%	86.14%

Table 35: Evaluation results on the patient classification of the proposed computer-aided CNN-CMGSs using the five examined CNN networks on the combined low and high magnification images of the original imbalanced JELEN_MERGE02 dataset. Best results indicated in bold.

As we mentioned before, the other comparison has been done among the examined CNN networks versus the data augmentation and rebalancing data techniques that were used to adjust the class distribution to improve the performance ability of the proposed CMGSs using the JELEN_MERGE02 dataset. To begin with, for data augmentation results presented in Tables 36, the GoogleNet-Inception-v3 network performs better than the other networks, reaching the best overall accuracy and sensitivity, while the ResNet50 network achieved the best specificity and precision for image classification. In Table 37, for patient classification, once again, the GoogleNet-Inception-v3 network reached the best overall accuracy and sensitivity while the ResNet101 network achieved the best specificity and precision.

CNN models	**Average accuracy**	**Average sensitivity**	**Average specificity**	**Average precision**
Inception-v3	**88.41**%	**95.47**%	57.26%	90.82%
ResNet101	87.05%	92.88%	61.32%	91.42%
ResNet50	86.56%	92.09%	**62.16**%	**91.50**%
ResNet18	85.90%	94.25%	49.05%	89.10 %
AlexNet	83.72%	95.14%	33.27%	86.38%

Table 36: Evaluation results on image classification for the proposed CNN-CMGSs using the five considered CNN networks on the combined low and high magnification images of the augmented JELEN_MERGE02 dataset. Best results indicated in bold.

On the other hand, for data sampling results, two data sampling techniques were performed as mentioned above, to begin with, data undersampling results are shown in Tables 38 and 39 using the combined low and high magnification images of the undersampled JELEN_MERGE02 dataset. For image classification, the ResNet101 network achieved the best overall accuracy and sensitivity rates and the AlexNet network reached the best specificity and precision rates. Meanwhile, both the GoogleNet-Inception-v3 and AlexNet achieved the best specificity rates, and again, AlexNet reached the best precision. In Table 39, for patient classification, the AlexNet network reached the best overall accuracy, specificity, and precision rates while the GoogleNet-Inception-v3 and ResNet101 networks both reached the best sensitivity.

For data oversampling results, in Table 40, the GoogleNet-Inception-v3 network achieved the best overall accuracy, sensitivity, specificity, and precision rates for image classification, whereas in Table 41, for patient classification, again, the GoogleNet-Inception-v3 network reached the best overall

CNN models	Average accuracy	Average sensitivity	Average specificity	Average precision
Inception-v3	**88.84**%	**99.87**%	47.72%	87.72%
ResNet101	87.98%	98.17%	**50.00**%	**88.02**%
ResNet50	86.82%	97.19%	48.18%	87.56%
ResNet18	84.71%	98.53%	33.18%	84.61%
AlexNet	81.34%	99.30%	41.39%	81.28%

Table 37: Evaluation results on patient classification for the proposed computer-aided CNN-CMGSs using the five CNN networks on the combined low and high magnification images of the augmented JELEN_MERGE02 dataset. Best results indicated in bold.

CNN models	Average accuracy	Average sensitivity	Average specificity	Average precision
ResNet101	**91.03**%	**93.01**%	89.05%	89.49%
Inception-v3	90.51%	90.18%	**90.84**%	90.89%
AlexNet	90.37%	89.90%	**90.84**%	**90.94**%
ResNet50	89.85%	91.41%	88.30%	88.70%
ResNet18	89.05%	91.41% 86.69%	86.69%	87.32%

Table 38: Evaluation results on image classification for the proposed computer-aided CNN-CMGSs using the five examined CNN networks on the combined low and high magnification images of the undersampled JELEN_MERGE02 dataset. Best results indicated in bold.

CNN models	Average accuracy	Average sensitivity	Average specificity	Average precision
AlexNet	**94.83**%	95.74%	**94.09**%	**93.11**%
Inception-v3	94.00%	**96.11**%	92.27%	91.24%
ResNet101	93.00%	**96.11**%	90.45%	89.25%
ResNet50	92.00%	95.00%	89.54%	88.27%
ResNet18	90.25%	95.55%	85.90%	84.90%

Table 39: Evaluation results on patient classification for the proposed computer-aided CNN-CMGSs using the five examined CNN networks on the combined low and high magnification images of the undersampled JELEN_MERGE02 dataset. Best results indicated in bold.

accuracy and sensitivity rates. The ResNet101 network achieved the best specificity and precision rates at 100%.

CNN models	Average accuracy	Average sensitivity	Average specificity	Average precision
Inception-v3	**96.36**%	**94.12**%	**98.61**%	**98.55**%
AlexNet	94.68%	91.98%	97.38%	97.24%
ResNet18	93.64%	89.40%	97.88%	97.70 %
ResNet50	92.70%	87.32%	98.07%	97.86%
ResNet101	92.62%	87.22%	98.03%	97.80%

Table 40: Evaluation results on image classification for the proposed computer-aided CNN-CMGSs using the five examined CNN networks based on the combined low and high magnification images of the oversampled JELEN_MERGE02 dataset. Best results indicated in bold.

According to the CNN results per computer-aided CMGS, it is readily apparent that the GoogleNet-Inception-v3 network nearly always has the statistically significant best performance with the two exceptions of ResNet101 (best for image classification) and AlexNet (best for patient classification) networks with the data undersampling approach. Based on these results, for the remaining experimental results, we will only consider the GoogleNet-Inception-v3 network. Thus, with the rest of the classification tasks (image and patient), we will expand our comparison to include the results

CNN models	Average accuracy	Average sensitivity	Average specificity	Average precision
Inception-v3	**99.71**%	**99.75**%	99.54%	99.87%
AlexNet	98.17%	97.80%	99.87%	99.87%
ResNet18	96.73%	95.97%	99.54%	99.87%
ResNet50	96.34%	95.48%	99.54%	99.87%
ResNet101	95.00%	93.65%	**100**%	**100**%

Table 41: Evaluation results on patient classification for the proposed computer-aided CNN-CMGSs using the five examined CNN networks on the combined low and high magnification images of the oversampled JELEN_MERGE02 dataset. Best results indicated in bold.

of the GoogleNet-Inception-v3 per image magnification and data type (imbalanced, augmented, undersampled, and oversampled). For the second pattern of classification where we used only the high magnification images to classify a certain case, in Tables 42 and 43 we display the results of the four evaluation methods that we obtained from the GoogleNet-Inception-v3 network using only the high magnification images that belong to each case. For image classification, in Table 42, the oversampled dataset reached the best overall accuracy, specificity, and precision rates, whereas the the imbalanced dataset achieved the best sensitivity. For patient classification, in table 43, again, the oversampled dataset reached the best overall accuracy, specificity, and precision rates, while the imbalanced dataset achieved the best sensitivity.

Data type	Average accuracy	Average sensitivity	Average specificity	Average precision
Imbalanced data	84.89%	**94.33**%	43.39%	88.12%
Augmented data	83.07%	92.87%	40.00%	87.26%
Undersampled data	80.66%	81.13%	80.18%	81.05%
Oversampled data	**94.14**%	90.47%	**97.81**%	**97.66**%

Table 42: Evaluation results on image classification for the proposed computer-aided CNN-CMGSs based on the GoogleNet-Inception-v3 network on only the high magnification images of JELEN_MERGE02 dataset. Best results indicated in bold.

Data type	Average accuracy	Average sensitivity	Average specificity	Average precision
Imbalanced data	85.96%	**98.17**%	40.45%	86.16%
Augmented data	82.50%	96.21%	31.36%	84.04%
Undersampled data	82.65%	87.40%	76.81%	82.73%
Oversampled data	**96.25**%	95.60%	**98.63**%	**99.62**%

Table 43: Evaluation results on patient classification for the proposed computer-aided CNN-CMGSs based on the GoogleNet-Inception-v3 network on only the high magnification images of JELEN_MERGE02 dataset. Best results indicated in bold.

In Tables 44 and 45, for the third pattern of image classification where we only considered the low magnification images to classify a certain case, we display the results of the four evaluation methods that we achieved from the GoogleNet-Inception-v3 network using only the low magnification images that belong to each case. For image classification, in table 44, the oversampled dataset reached the best overall accuracy and specificity rates, whereas the the imbalanced dataset achieved the best specificity and precision rates. For the patient classification, in table 45, the oversampled dataset reached the best overall accuracy, specificity, and precision rates, while the imbalanced dataset achieved the best sensitivity of 100%.

Data type	Average accuracy	Average sensitivity	Average specificity	Average precision
Imbalanced data	89.98%	**96.21**%	62.45%	**91.89**%
Augmented data	87.18%	94.46%	54.90%	90.30%
Undersampled data	82.92%	84.33%	81.50%	82.49%
Oversampled data	**92.50**%	93.53%	**88.63**%	90.13%

Table 44: Evaluation results on image classification for the proposed computer-aided CNN-CMGSs based on the GoogleNet-Inception-v3 network on only the low magnification images of JELEN_MERGE02 dataset. Best results indicated in bold.

As we mentioned before in the classification section, primitive attempts were performed to classify each case (pair of images) that correspond to each patient as belonging to malignancy grade G2 or

Data type	Average accuracy	Average sensitivity	Average specificity	Average precision
Imbalanced data	90.00%	**100.00**%	52.72%	88.81%
Augmented data	87.78%	98.04%	49.54%	87.90%
Undersampled data	84.48%	90.00%	77.72%	83.64%
Oversampled data	**96.25**%	95.60%	**98.63**%	**99.62**%

Table 45: Evaluation results on patient classification for the proposed computer-aided CNN-CMGSs based on the GoogleNet-Inception-v3 network on only the low magnification images of JELEN_MERGE02 dataset. Best results indicated in bold.

G3. In Tables 46 and 47, we display the results of the four evaluation methods that we obtained from the GoogleNet-Inception-v3 network using the concatenated pairs of images that belong to each patient. For the case classification, in table 46, the oversampled dataset reached the best overall accuracy of 84.61%, whereas the augmented data achieved the best sensitivity and specificity, while the imbalanced dataset achieved the best precision rates. For patient classification, in table 47, the oversampled dataset reached the best overall accuracy, specificity, and precision rates, while the augmented dataset achieved the best sensitivity.

Data type	Average accuracy	Average sensitivity	Average specificity	Average precision
Imbalanced data	81.11%	90.42%	40.18%	**87.05**%
Augmented data	81.15%	**90.55**%	**86.99**%	86.99%
Undersampled data	76.41%	78.67%	74.15%	76.17%
Oversampled data	**84.61**%	84.16%	85.06%	85.21%

Table 46: Evaluation results on case classification for the proposed computer-aided CNN-CMGSs based on the GoogleNet-Inception-v3 network on the concatenated images of each pair of JELEN_MERGE02 dataset. Best results indicated in bold.

5.3.1 Comparison with State-of-the-Art

A direct comparison with previous work, based on handcrafted features and traditional classifiers such as SVM, is complicated by the fact that in this previous work case classification was considered

Data type	Average accuracy	Average sensitivity	Average specificity	Average precision
Imbalanced data	81.34%	94.39%	32.72%	84.11%
Augmented data	82.01%	**94.75%**	34.54%	84.56%
Undersampled data	81.16%	86.19%	76.36%	79.01%
Oversampled data	**90.48%**	90.12%	**91.81%**	**97.61%**

Table 47: Evaluation results on patient classification for the proposed computer-aided CNN-CMGSs based on the GoogleNet-Inception-v3 network on the concatenated images of each pair of JELEN_MERGE02 dataset. Best results indicated in bold.

whereas in this current work image classification is considered. Nonetheless, in Figure **??**, we give a side by side comparison table. The main comparison that can me made is with patient classification. As can clearly be seen, in terms of average accuracy, the current work achieves more than 3% improvement over previous work, while sensitivity, specificity and precision are all improved as well.

Cytological Grading Systems	Case Classification				Patient Classification			
	Accuracy	Sensitivity	Specificity	Precision	Accuracy	Sensitivity	Specificity	Precision
Khan *et al.* CA-CGS [5]	96.66%	88.68%	98.64%	94.24%	96.24%	89.33%	98.40%	94.66%
Robinson CA-CGS [5]	97.28%	93.33%	98.27%	93.09%	95.29%	85.77%	98.26%	93.99%
CNN models	**Image Classification**				**Patient Classification**			
	Accuracy	Sensitivity	Specificity	Precision	Accuracy	Sensitivity	Specificity	Precision
Inception-v3 CNN-CMGS	96.36%	98.61%	94.12%	94.4%	99.71%	99.54%	99.75%	99.13%

Figure 61: Comparison of CA-CGSs from chapter 3 and the best result from this chapter on the JELEN_MERGE02 dataset .

5.4 Discussion and conclusions

The motivation behind using DL was to get rid of the images preprocessing task for the classification of malignancy grading problem. However, this technique comes with its own challenges. One of the challenges faced was again related to the imbalanced data problem which strongly impacted the specificity performance of the used CNN. Specifically, although GoogleNet-Inception-v3 had achieved accuracy close to 90% for both image and patient classification, the specificity in both the case and patient were close to 60% and the tendency for the specificity to decrease for imbalanced data was in agreement with the study of handcrafted based feature systems (low performance from the CNN models on the minority class prediction that represent the G3 samples). This means that

imbalanced data is a significant problem for both ML and CNN systems as they become highly biased towards the majority class and less sensitive to the minority class.

According to the literature [138, 139, 140], to solve the problem of imbalanced data, a generation of synthetic samples using data augmentation was usually performed for image classification based on CNN models, specifically for limited or imbalanced data. However, the results obtained were not in line with the related literature and showed the same trend as the results networks dealing with imbalanced data. This happened because data augmentation did not improve the balancing of the data but instead increased the number of data in the same proportion for both the majority and minority class, which implies that the ratio of the classes remains the same. This urged us to further investigate the rebalancing of the dataset as was performed for handcrafted based CA-CGSs.

The techniques applied for rebalancing were undersampling and oversampling. With regard to undersampling, the best results were obtained for ResNet-101 with an accuracy of 91.03% for image classification and specificity close to 90%, illustrating that the network is trained for handling minority classes, thereby reducing the bias of the network. Meanwhile, for patient classification, AlexNet gave the highest accuracy performance of 94.83% with a specificity close to 95%.

Though the specificity was improved for all the networks, the highest value achieved was 94.09%, which means there is still almost a 6% error for detecting the minority class, G3. Thus, aiming to improve this high error rate, the oversampling method was applied which resulted in an improvement of the accuracy of the system as well as an increase of the specificity to 98.61% by using GoogleNet-Inception-v3 for image classification. For patient classification, the specificity rose to 100% for ResNet101 but the accuracy using ResNet101 was only 95%, which is an indication of overfitting in the minority class. Thus, GoogleNet-Inception-v3 was preferred over all the networks with an accuracy of 99.71% and a specificity rate of 99.54%, thereby solving both the overfitting and rebalancing problems. Indeed, applying the oversampling technique greatly decreased the biases which consequently improved the specificity by approximately 40%. Thus, it is highly recommended to balance the data before using any CNN model.

Typically, the results from both magnifications or only high magnification are preferred for predicting the malignancy grading of CGSs. Only low magnification images are typically ignored by doctors for the grading of cancer. Therefore, in this research, special attention was given to low magnification images to classify the malignancy grading of CGSs. A total of 288 low magnification images out of which 235 G2 and 53 G3 images were classified using GoogleNet-Inception-v3 by applying the oversampling technique. An accuracy of 92.50% and 96.25% was achieved with a specificity of 88.63% and 98.63% for image and patient classification, respectively. These results showed that low

magnification also plays an important role in determining the malignancy grade of cancer, and the results obtained are highly correlated with the previously modified CA-CGSs.

After considering image and patient classification, case classification was taken into account by concatenating high and low magnification image which considered both the magnification features instead of two separate images. The advantage was to consider the case classification problem which was not possible from the previous DL methods. Again, GoogleNet-Inception-v3 was preferred with the oversampling technique to give an accuracy of 84.61% and 90.48% for the case and patient classification, respectively. The accuracy for this classification is low as compared to the previous results which might be due to the loss of information from the concatenation and resizing process. This process is a proof of concept and a step forward for the future work which will concatenate the results from the separated low and high magnification networks, and provides the final case classification result without losing any information.

Chapter 6

Discussion and conclusions

In this chapter, a summary of the main contributions of this thesis and the possible impact as we noticed from the possible impact of our research is discussed. Further, limitations, and important future work directions are reported.

6.1 Summary of thesis contributions

1. **Proposed nine computer-aided CGSs for FNA biopsies of breast cancer:** Due to the alarming number of yearly breast cancer-related deaths in middle-aged women, automatic diagnosis and grading of cancer cells have become one of the major research areas for both medical image classification scientists and clinicians. Three main objectives to accurately automate the malignancy grading problem for FNA biopsies of breast cancer were considered in this thesis. To begin with, six computer-aided CGSs for FNAs of breast cancer based on six published cytological malignancy grading schemes used by pathologists (Robinson's CGS; Khan *et al.*'s CGS; Fisher's CGS; etc.) were proposed as well as three novel CA-CGSs based on modifications to the original versions of Fisher's, Khan *et al.*'s and Taniguchi *et al.*'s systems by adding the low magnification features to the original systems. To maintain a strong connection between the decision-making process of the pathologist using one of these CGSs and the determination of the computer-aided CGSs, in our classification systems, the features were estimated and used to best express the cytological characteristics used by pathologists. The following is a summary of the main contributions of this thesis.

    - **Cytological image segmentation task:** One of the most important steps in the classification of cytological images is the pre-processing of the images in order to aid machine

learning algorithms in the accurate classification of images. First, color deconvolution was performed using optical density matrix. This step was done to separate color information from cytoplasm and nuclei regions in order to improve the image segmentation results. Next, image contrast enhancement, another pre-processing step, was conducted using histogram stretching to adjust image intensity values which increase the contrast of the output images. This step is essential for the obtainment of the initial enhanced boundary after applying multi-level thresholds by Otsu's segmentation method. Then, the quantization process was done on the basis of the calculated multi-level thresholds to segment the images into three regions distinguished by three different colors. It was thereafter converted to new, enhanced RGB images to extract the channel of interest. It is highlighted here that all these steps were performed in order to get a good initial mask to feed into the image segmentation algorithm where the GVF-snake algorithm was used due to its ability to detect each and every pixel of the boundaries of the interesting object. It is a parametric algorithm which requires an initial contour, thus the green boundaries achieved via the previous steps were supplied as an initial contour to segment nuclei regions, whereas blue boundaries were used to segment cytoplasm regions. Although the GVF-MO gave good segmentation results, due to different sizes of the nuclei region in the images, the algorithm was not able to separate some of the clusters of the nuclear regions and considered them to be one big nucleus. This was undesirable since the aim was to detect each nuclei region for the feature extraction purpose. To solve this issue, the watershed algorithm was applied to the top of this segmented result. The most interesting feature of this algorithm is its ability to segment the cluster of the connected nuclei based on the given diameter of the nuclei region. In this work, the size of the nuclei region was determined numerically to be 100 pixels. This means that if the size of the object is greater than the given diameter, then it is considered as a cluster of connected nuclei and is re-segmented to individual nuclei. In the final segmented images obtained, there were nuclei regions, and also some unwanted regions such as red blood cells overlapping nuclei, and arrangements in the background, treated as nuclei regions. In the final segmentation step, nuclei filtration was applied according to seven numerically estimated features using the SVM classifier with manually labeled training data. The accuracy achieved was 80.24%, evaluated using 5-fold cross-validation technique. The whole preprocess ensured that only nuclei regions were used for the feature extraction task while bad nuclei regions were ignored. The robustness of the segmentation was evaluated based on the classification accuracy results. Typically, precise image segmentation results yield accurate feature extraction which in turn produce accurate classification results (was evaluated based on

the ground truth labels).

- **Handcrafted feature extraction task:** Following the segmentation, the feature extraction stage was implemented using the considered six CGSs where each CGS has its own cytological criteria to determine the malignancy grade of cancer. These criteria were mimicked in this thesis to determine the malignancy grade of cancer. In this context, there were challenges related to the estimation of some of these features. The first challenge was the mitosis estimation because the FNA images were not made for this purpose. To solve this issue, the mitosis was estimated using nuclei division instead of cells based on 25 morphological and textural features. This was possible because a nuclei division reflects a cell division. The second challenge was to estimate CN (a type of cell death) because during the image preparation cell death can occur naturally due to material extraction and the staining procedure. Thus, it was not possible for FNA cytological images to estimate whether cell death occurred due to a disease or not. The third challenge was to estimate naked tumor nuclei which becomes difficult due to the significant variations in the sizes of the nuclei. To solve this issue, the ratios of the number of very big nuclei and very small nuclei to the total number of nuclei in each image were calculated according to a threshold value experimentally determined based on nuclei size criteria, where it naked nucleus was assumed if exceeding that threshold value. On the other hand, because the cell structure is not preserved for cytological images, the fourth challenge was related to the cellular characteristics. Instead, nuclear features were used to determine cellular characteristics. For example, if cellular uniformity was to be estimated then it was correlated with the nuclear uniformity due to the fact that any change in the cell starts from its heart, the nucleus.

- **Feature selection and classification tasks:** For the feature selection purpose, a correlation score between the features and the labeled output was calculated using the Fisher feature selection algorithm. The top 30% of the features were selected as the optimal subset for the classification purpose. For the classification task, two classification schemes were performed: the case (each case classified individually) and patient classification (each patient classified based on taking majority voting of his cases classification results). Further, nine different classifiers were used in both classification schemes to classify the malignancy grade of cancer. It was found that out of the nine used classification algorithms the SVM performed best for both case and patient classification for most of the proposed CGSs.

2. **Rebalancing the class distribution:** The results obtained from the classification algorithm

systems were satisfactory; however, there was a significant problem with data imbalance. The performance of existing traditional learning algorithms can be affected by different learning aspects. An imbalanced dataset is a dataset that contains more samples belonging to one class (so-called majority class) compared to the other classes (namely minority classes). In class imbalanced classifications, classifiers can usually achieve more sensitive rates towards the majority class and low sensitive rates towards the minority class which leads to poor performance accuracy on the minority class in comparison to the majority class. Thus, in this work, three different data balancing techniques, namely, oversampling, undersampling, and, Hybrid RUSBoost, were considered to increase the sensitivity performance in the minority class. The results show a significant improvement in the sensitivity and precision of the classifiers compared to the imbalanced data classification case. This means that the resulting classifiers are no longer biased towards one class and will perform better irrespective of the given class. One of the side effects of the oversampling technique was overfitting to the training data due to which the accuracy was compromised by a slight margin.

3. **Introducing the deep learning approach for the problem of breast cancer grading:** Although useful, it is noted that the complete process of handcrafted features-based CGSs is time-dependent and very subjective to the individual's experience. Therefore, there was a need to introduce a system which is less subjective and less time-consuming. To this end, DL was introduced to facilitate all the above mentioned manual tasks by means of an automated and equally accurate approach. According to the results obtained, in terms of the computer-aided grading system per CNN model, the pre-trained GoogleNet-inception-v3 network outperformed the other examined pre-trained networks for both image and patient classification results. The pre-trained network was used because, originally, the data was very limited and deep learning requires a lot of data. Since the pre-trained network was pre-trained on natural images and which differ from medical images, the accuracy was slightly compromised. The other challenge was related to the imbalanced data which impacts the sensitivity of the data. This was solved using the oversampling method, which achieved the best accuracy as compared to other datasets and has significantly improved the performance of the proposed CNN-based CMGSs compared to the case with the original imbalanced dataset.

4. **Results evaluation** The overall obtained results conclude that CGSs have high correlation among each other since they share some of the malignant criteria and their accuracy are very close to one another. Thus, doctors can use any one of them except Howell's system. This may be due to the lack of inclusion of more nuclear characteristics. These results demonstrate that computer-aided breast cancer malignancy grading systems using FNA might be able to achieve

accuracy rates comparable to its more invasive histopathological BR system counterpart.

6.2 Overall evaluation results for the proposed grading systems

For the traditional classification system, Robinson's-based CA-CGS achieved the best results for the case classification problem with (97%) overall accuracy, outperforming the accuracy of the pathology-based studies (90%) and the previous work by Bruździński et al (89.02%).

For patient classification problem, Khan et al.'s-based CGS achieved the best accuracy of (96%), which is a trustworthy result since it is the best with imbalanced data (96.50%) and rebalanced data (99.61%).

Ultimately, the modified version of Khan et al.'s was determined to be the best for both case and patient classification tasks (x percentage and y percentage, respectively).

For the CNN-based grading system, Inception-V3 CA-CMGS was determined to be the the best with 99.71

Finally, we recommend the modified Khan et al.'s CA-CGS and Inception-V3 CNN-CMG systems to be used in real medical environments, as they are not only the most accurate systems, but also save time, costs, and most importantly, save lives.

6.3 Open problem and future work

This thesis sought to automatize the cytological malignancy grading task for FNA biopsies of breast cancer. The aim was to precisely distinguish between the classified samples as belonging to malignancy grade G2 or G3, where a variety of image processing and machine learning algorithms as well as deep learning models were used for this purpose. The research topics covered in the dissertation provide multiple possible extensions. The following is a summary of possible future work.

To date, there are several open points that can be summarized here:

1. **Dataset expansion and consistency:** Collecting more training dataset from both case and patient samples will improve the overall accuracy performance due to the fact that adding more training dataset will decrease the overfitting problem. Further, the consistency in the number of cases that belong to each patient will enhance the accuracy performance for the patient classification.

2. **Expand low magnification features:** Due to the discriminatory power of the low magnification features on boosting the overall performance accuracy of the grading systems, it will be worthy to expand these features.

3. **Maintain direct cooperation with medical centers:** Because of the challenging work and some limitations during feature extraction to simulate variety sets of features to reflect the pathology-based criteria of the grading systems, it is substantial to provide direct cooperation with breast cancer medical centers to overcome the limitations associated with the estimation of the malignant criteria.

4. **Collect a ground truth for the breast cancer datasets:** Due to the mentioned challenges and limitations associated with the feature extraction, it is essential to provide ground truth information of the used breast cancer dataset such as the normal nuclei size, labeled MC and labeled CN features.

5. **CNN-based case classification model:** Implementation of the convolutional neural networks for the case classification problem which was not systematically considered in this thesis.

6. **Collect low malignancy grade G1 samples:** Collection of low malignancy grade G1 samples to develop a comprehensive cytological grading for FNA biopsies for breast cancer that could provide great assistance to breast cancer medical centers.

7. **Single cell analysis:** It will be more worthwhile to consider the single cell analysis by investigating the cell segmentation task and cellular morphological and polymorphic feature extraction. This will improve the overall performance of the cytological grading systems since it will consider the cell's evaluation, and not only the nuclear evaluation.

Bibliography

[1] Rebecca L. Siegel, Kimberly D. Miller, and Ahmedin Jemal. Cancer statistics, 2017. *CA: A Cancer Journal for Clinicians*, 67(1):7–30, 2017.

[2] International Agency for Research on Cancer et al. Latest world cancer statistics global cancer burden rises to 14.1 million new cases in 2012: Marked increase in breast cancers must be addressed. *World Health Organization*, 12, 2013.

[3] R Alteri et al. Cancer facts & figures 2013. *Atlanta: American Cancer Society*, 2013.

[4] Suvradeep Mitra and Pranab Dey. Fine-needle aspiration and core biopsy in the diagnosis of breast lesions: A comparison and review of the literature. *Cytojournal*, 13, 2016.

[5] Clifton D Bryant. *Handbook of death and dying*, volume 1. Sage, 2003.

[6] Stephen S Raab, Dana Marie Grzybicki, Janine E Janosky, Richard J Zarbo, Frederick A Meier, Chris Jensen, and Stanley J Geyer. Clinical impact and frequency of anatomic pathology errors in cancer diagnoses. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, 104(10):2205–2213, 2005.

[7] Chrisoula D Scopa, Demetrios Koukouras, John Androulakis, and Dionysis Bonikos. Sources of diagnostic discrepancies in fine-needle aspiration of the breast. *Diagnostic Cytopathology*, 7(5):546–548, 1991.

[8] Inc. American Cancer Society. Surgery for breast cancer. `https://www.cancer.org/cancer/breast-cancer/treatment/surgery-for-breast-cancer.html`, 2019. [Online; accessed 2019].

[9] Shelley McGuire. World cancer report 2014. geneva, switzerland: World health organization, international agency for research on cancer, who press, 2015, 2016.

[10] Lukasz Jeleri. *Computerized cancer malignancy grading of fine needle aspirates.* PhD thesis, Citeseer, 2009.

[11] Chimezie I Madubogwu, Cornelius O Ukah, Igwebuike V Onyiaorah, Daniel CD Anyiam, Stanley NC Anyanwu, and Gabriel U Chianakwana. Cost effectiveness of fine needle aspiration cytology for breast masses. *Orient Journal of Medicine*, 27(1-2):22–27, 2015.

[12] US Susan G Komen Breast Cancer Foundation, Inc. Fine needle aspiration (fine needle biopsy). `https://ww5.komen.org/BreastCancer/FineNeedleBiopsy.html`, 2019. [Online; accessed 2019].

[13] National national institutes of health. `http://www.cancer.gov/types/breast/hp/breast-screening-pdq#cit/section_1.13/`, note = Online; accessed 2016-08-21.

[14] Svante R Orell and Gelareh Farshid. False-positive reports in fine needle biopsy of breast lesions. *Pathology*, 33(4):428–436, 2001.

[15] MZ Rahman, AM Sikder, and SR Nabi. Diagnosis of breast lump by fine needle aspiration cytology and mammography. *Mymensingh Medical Journal: MMJ*, 20(4):658–664, 2011.

[16] Yun Gong. Breast cancer: Pathology, cytology, and core needle biopsy methods for diagnosis. In *Breast and Gynecological Cancers*, pages 19–37. Springer, 2013.

[17] Inc. National Breast Cancer Foundation. Janelle hail. `http://www.nationalbreastcancer.org/breast-cancer-biopsy/`, 2015. [Online; accessed 2015].

[18] National Cancer Institute. Nci. `http://www.cancer.gov/about-cancer/diagnosis-staging/prognosis/tumor-grade-fact-sheet/`, 2015.

[19] HJG Bloom and WW Richardson. Histological grading and prognosis in breast cancer: a study of 1409 cases of which 359 have been followed for 15 years. *British Journal of Cancer*, 11(3):359, 1957.

[20] D.G. Hicks and S.C. Lester. *Diagnostic Pathology: Breast.* Diagnostic pathology. Amirsys, 2011.

[21] Qicheng Lao and Thomas Fevens. Case-based histopathological malignancy diagnosis using convolutional neural networks. In *BMVC*, 2017.

[22] JL Haybittle, RW Blamey, CW Elston, Jane Johnson, PJ Doyle, FC Campbell, RI Nicholson, and K Griffiths. A prognostic index in primary breast cancer. *British Journal of Cancer*, 45(3):361, 1982.

[23] Vidya Vasudev, R Rangaswamy, and V Geethamani. The cytological grading of malignant neoplasms of the breast and its correlation with the histological grading. *Journal of Clinical and Diagnostic Research: JCDR*, 7(6):1035, 2013.

[24] Christopher W Elston and Ian O Ellis. Pathological prognostic factors in breast cancer. i. the value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology*, 19(5):403–410, 1991.

[25] Sangita Bhattacharjee, Jashojit Mukherjee, Sanjay Nag, Indra Kanta Maitra, and Samir K Bandyopadhyay. Review on histopathological slide analysis using digital microscopy. *International Journal of Advanced Science and Technology*, 62:65–96, 2014.

[26] Lei He, L Rodney Long, Sameer Antani, and George R Thoma. Histology image analysis for carcinoma detection and grading. *Computer Methods and Programs in Biomedicine*, 107(3):538–556, 2012.

[27] Kasonde Bowa, Jim Jewel, and Victor Mudenda. Fine needle aspiration cytology in the investigation of breast lumps at the university teaching hospital in lusaka, zambia. *Tropical Doctor*, 38(4):245–247, 2008.

[28] Joseph Ragaz and Irving M Ariel. *High-risk breast cancer: diagnosis*. Springer Science & Business Media, 2012.

[29] Cecilia Bozzetti, Rita Nizzoli, Nadia Naldi, Annamaria Guazzi, Roberta Camisa, Laura Manotti, Francesco Paolo Pilato, Giuliano Mazzini, and Giorgio Cocconi. Nuclear grading and flow cytometric dna pattern in fine-needle aspirates of primary breast cancer. *Diagnostic Cytopathology*, 15(2):116–120, 1996.

[30] C. Bansal, U.S. Singh, S. Misra, K.L. Sharma, V. Tiwari, and A.N. Srivastava. Comparative evaluation of the modified Scarff-Bloom-Richardson grading system on breast carcinoma aspirates and histopathology. *Cytojournal*, 9:4, January 2012. [Online Journal].

[31] Houghton Mifflin Harcourt. The cell and its membrane. `https://www.cliffsnotes.com/study-guides/anatomy-and-physiology/the-cell/the-cell-and-its-membrane`, 2016. [Online; accessed 2019].

[32] Pinki Pandey, Alok Dixit, Subrat Chandra, and Swarn Kaur. A comparative and evaluative study of two cytological grading systems in breast carcinoma with histological grading: an important prognostic factor. *Analytical Cellular Pathology*, 2014, 2014.

[33] AK Das, Kusum Kapila, AK Dinda, and Kusum Verma. Comparative evaluation of grading of breast carcinomas in fine needle aspirates by two methods. *Indian Journal of Medical Research*, 118:247, 2003.

[34] E.R. Fisher, C. Redmond, and B. Fisher. Histologic grading of breast cancer. *Pathology Annual*, 15(Pt 1):239–251, 1980.

[35] J. Mouriquand, M. Gozlan-Fior, D. Villemain, Y. Bouchet, J.C. Sage, M.A. Mermet, and M. Bolla. Value of cytoprognostic classification in breast carcinomas. *Journal of Clinical Pathology*, 39:489—496, 1986.

[36] I.A. Robinson, G. McKee, A. Nicholson, P.A. Jackson, M.G. Cook, J. D'Arcy, and M.W. Kissin. Prognostic value of cytological grading of fine-needle aspirates from breast carcinomas. *The Lancet*, 343(8903):947–949, 1994. Originally published as Volume 1, Issue 8903.

[37] L.P. Howell, R. Gandour-Edwards, and D. O'Sullivan. Application of the Scarff-Bloom-Richardson tumor grading system to fine-needle aspirates of the breast. *American Journal of Clinical Pathology*, 101:262–265, 1994.

[38] M.Z. Khan, A. Haleem, H. Al Hassani, and H. Kfoury. Cytopathological grading, as a predictor of histopathological grade, in ductal carcinoma (NOS) of breast, on air-dried Diff-Quik smears. *Diagnostic Cytopathology*, 29:185–193, 2003.

[39] E. Taniguchi, Q. Yang, W. Tang, Y. Nakamura, L. Shan, M. Nakamura, M. Sato, I. Mori, T. Sakurai, and K. Kakudo. Cytologic grading of invasive breast carcinoma. Correlation with clinicopathologic variables and predictive value of nodal metastasis. *Acta Cytologica*, 44:587–591, 2000.

[40] Kaushik Saha, Gargi Raychaudhuri, Bitan Kuamr Chattopadhyay, and Indranil Das. Comparative evaluation of six cytological grading systems in breast carcinoma. *Journal of Cytology/Indian Academy of Cytologists*, 30(2):87, 2013.

[41] Dinisha Einstien, BO Omprakash, Hemalatha Ganapathy, and Sadaf Rahman. Comparison of 3-tier cytological grading systems for breast carcinoma. *Isrn Oncology*, 2014, 2014.

[42] Maurice M Black. Survival in breast cancer cases in relation to the structure of the primary tumor and regional lymphnodes. *Surg Gynecol Obstet*, 100:543–551, 1955.

[43] EDWIN R Fisher, CAROL Redmond, and BERNARD Fisher. Histologic grading of breast cancer. *Pathology Annual*, 15(Pt 1):239, 1980.

[44] Kyoung-Mi Lee and W Nick Street. An adaptive resource-allocating network for automated detection, segmentation, and classification of breast cancer nuclei topic area: image processing and recognition. *IEEE Transactions on Neural Networks*, 14(3):680–687, 2003.

[45] Cigdem Demir and Bülent Yener. Automated cancer diagnosis based on histopathological images: a systematic survey. *Rensselaer Polytechnic Institute, Tech. REP*, 2005.

[46] John B Zimmerman, Stephen M Pizer, Edward V Staab, J Randolph Perry, William McCartney, and Bradley C Brenton. An evaluation of the effectiveness of adaptive histogram equalization for contrast enhancement. *IEEE Transactions on Medical Imaging*, 7(4):304–312, 1988.

[47] Rafael C Gonzalez and E Richard. Woods, digital image processing. *ED: Prentice Hall Press, ISBN 0-201-18075-8*, 2002.

[48] Erik Meijering. Cell segmentation: 50 years down the road [life sciences]. *IEEE Signal Processing Magazine*, 29(5):140–145, 2012.

[49] Mehmet Sezgin et al. Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic Imaging*, 13(1):146–168, 2004.

[50] Łukasz Jeleń, Adam Krzyżak, and Thomas Fevens. Comparison of pleomorphic and structural features used for breast cancer malignancy classification. In *Advances in Artificial Intelligence*, pages 138–149. Springer, 2008.

[51] Ezzatollah Salari and Pepe Siy. The ridge-seeking method for obtaining the skeleton of digital images. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, (3):524–528, 1984.

[52] Pawel Filipczuk, Thomas Fevens, Adam Krzyzak, and Roman Monczak. Computer-aided breast cancer diagnosis based on the analysis of cytological images of fine needle biopsies. *IEEE Transactions on Medical Imaging*, 32(12):2169–2178, 2013.

[53] Lukasz Jelen, Thomas Fevens, and Adam Krzyzak. Influence of nuclei segmentation on breast cancer malignancy classification. In *SPIE Medical Imaging*, pages 726014–726014. International Society for Optics and Photonics, 2009.

[54] Robert M Haralick, Karthikeyan Shanmugam, and Its' Hak Dinstein. Textural features for image classification. *SIEEE Transactions on Systems, Man, and Cybernetics: Systems*, (6):610–621, 1973.

[55] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (12):2037–2041, 2006.

[56] Ingrid Daubechies. Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, 41(7):909–996, 1988.

[57] Jason Brownlee. Title of citation. `http://machinelearningmastery.com/an-introduction-to-feature-selection/`, October 6, 2014.

[58] Mark A Hall. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999.

[59] S Vanaja and K Ramesh Kumar. Analysis of feature selection algorithms on classification: a survey. *International Journal of Computer Applications*, 96(17), 2014.

[60] Christopher M Bishop. Pattern recognition. *Machine Learning*, 128, 2006.

[61] George H John and Pat Langley. Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 338–345. Morgan Kaufmann Publishers Inc., 1995.

[62] Richard O Duda, Peter E Hart, and David G Stork. *Pattern Classification*. John Wiley & Sons, second edition, 2001.

[63] Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. An knn model-based approach and its application in text categorization. In *Computational Linguistics and Intelligent Text Processing*, pages 559–570. Springer, 2004.

[64] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[65] Joseph A Cruz and David S Wishart. Applications of machine learning in cancer prediction and prognosis. *Cancer Informatics*, 2:59, 2006.

[66] Jason Brownlee. How to configure the number of layers and nodes in a neural network, September 2018.

[67] Abraham Pouliakis, Efrossyni Karakitsou, Niki Margari, Panagiotis Bountris, Maria Haritou, John Panayiotides, Dimitrios Koutsouris, and Petros Karakitsos. Artificial neural networks as decision support tools in cytopathology: past, present, and future. *Biomedical Engineering and Computational Biology*, 7:BECB–S31601, 2016.

[68] Taha Eren Sarnıç. Estimating the religion of countries according to shapes of the flags using support vector machines and kernel methods. *Issues*, 1(1), 2014.

[69] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[70] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[71] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *Arxiv Preprint Arxiv:1502.03167*, 2015.

[72] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[73] Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization? In *Advances in Neural Information Processing Systems*, pages 2483–2493, 2018.

[74] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[75] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[76] Muhammad Rizwan. Residual networks (resnets). `https://engmrk.com/residual-networks-resnets/?utm_campaign=News&utm_medium=Community&utm_source=DataCamp.com`, 2019. [Online; accessed 2019].

[77] Ron Kohavi. Glossary of terms. *Special Issue on Applications of Machine Learning and The Knowledge Discovery Process*, 30(271):127–132, 1998.

[78] Metin N Gurcan, Laura E Boucheron, Ali Can, Anant Madabhushi, Nasir M Rajpoot, and Bulent Yener. Histopathological image analysis: A review. *IEEE Reviews in Biomedical Engineering*, 2:147–171, 2009.

[79] S Chhabra, PK Singh, A Agarwal, A Bhagoliwal, and SN Singh. Cytological grading of breast carcinoma: A multivariate regression analysis. *J Cytol*, 22(2):62–5, 2005.

[80] Ajay Basavanhally, Shannon Agner, Gabriela Alexe, Gyan Bhanot, Shridar Ganesan, and Anant Madabhushi. Manifold learning with graph-based features for identifying extent of lymphocytic infiltration from high grade her2+ breast cancer histology. In *Workshop on Microscopic Image Analysis with Applications in Biology*. Citeseer, 2008.

[81] Scott Doyle, Shannon Agner, Anant Madabhushi, Michael Feldman, and John Tomaszewski. Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features. In *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 496–499. IEEE, 2008.

[82] Shivang Naik, Scott Doyle, Shannon Agner, Anant Madabhushi, Michael Feldman, and John Tomaszewski. Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology. In *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 284–287. IEEE, 2008.

[83] Łukasz Jeleń, Thomas Fevens, and Adam Krzyżak. Classification of breast cancer malignancy using cytological images of fine needle aspiration biopsies. *International Journal of Applied Mathematics and Computer Science*, 18(1):75–83, 2008.

[84] Łukasz Jeleń, Adam Krzyżak, Thomas Fevens, and Michał Jeleń. Influence of feature set reduction on breast cancer malignancy classification of fine needle aspiration biopsies. *Computers in Biology and Medicine*, 79:80–91, 2016.

[85] Tomasz Bruździński, Adam Krzyżak, Thomas Fevens, and Łukasz Jeleń. Web–based framework for breast cancer classification. *Journal of Artificial Intelligence and Soft Computing Research*, 4(2):149–162, 2014.

[86] Łukasz Jeleń, Artur Lipiński, Jerzy Detyna, and Michał Jeleń. Grading breast cancer malignancy with neural networks. *Editorial Board*, page 47, 2011.

[87] Lukasz Jelen, T Fevens, A Krzyżak, and M Jeleń. Discriminatory power of cells grouping features for breast cancer malignancy classification. In *4th Kuala Lumpur International Conference on Biomedical Engineering 2008*, pages 559–562. Springer, 2008.

[88] Bartosz Krawczyk, Łukasz Jeleń, and Michał Woźniak. Adaptive splitting and selection ensemble for breast cancer malignancy grading. In *Computational Intelligence in Healthcare and e-health (CICARE), 2014 IEEE Symposium on*, pages 104–111. IEEE, 2014.

[89] Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM Sigkdd Explorations Newsletter*, 6(1):20–29, 2004.

[90] Bartosz Krawczyk, Mikel Galar, Łukasz Jeleń, and Francisco Herrera. Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy. *Applied Soft Computing*, 38:714–726, 2016.

[91] Maciej A Mazurowski, Piotr A Habas, Jacek M Zurada, Joseph Y Lo, Jay A Baker, and Georgia D Tourassi. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural Networks*, 21(2):427–436, 2008.

[92] Hongyu Guo and Herna L Viktor. Learning from imbalanced data sets with boosting and data generation: the databoost-im approach. *ACM Sigkdd Explorations Newsletter*, 6(1):30–39, 2004.

[93] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833. Springer, 2014.

[94] Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-Jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M Hoffman, et al. Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, 15(141):20170387, 2018.

[95] Diogo M Camacho, Katherine M Collins, Rani K Powers, James C Costello, and James J Collins. Next-generation machine learning for biological networks. *Cell*, 2018.

[96] Alexey A Shvets, Alexander Rakhlin, Alexandr A Kalinin, and Vladimir I Iglovikov. Automatic instrument segmentation in robot-assisted surgery using deep learning. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 624–628. IEEE, 2018.

[97] Babak Ehteshami Bejnordi, Maeve Mullooly, Ruth M Pfeiffer, Shaoqi Fan, Pamela M Vacek, Donald L Weaver, Sally Herschorn, Louise A Brinton, Bram van Ginneken, Nico Karssemeijer, et al. Using deep convolutional neural networks to identify and classify tumor-associated stroma in diagnostic breast biopsies. *Modern Pathology*, page 1, 2018.

[98] Alexander Rakhlin, Alexey Shvets, Vladimir Iglovikov, and Alexandr A Kalinin. Deep convolutional neural networks for breast cancer histology image analysis. In *International Conference Image Analysis and Recognition*, pages 737–744. Springer, 2018.

[99] Michał Żejmo, Marek Kowal, Józef Korbicz, and Roman Monczak. Classification of breast cancer cytological specimen using convolutional neural network. In *Journal of Physics: Conference Series*, volume 783, page 012060. IOP Publishing, 2017.

[100] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T. Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*, 19(6):1236–1246, 2018.

[101] Arnout C Ruifrok, Dennis A Johnston, et al. Quantification of histochemical staining by color deconvolution. *Analytical and Quantitative Cytology and Histology*, 23(4):291–299, 2001.

[102] Ahmed Elnakib, Georgy Gimel'farb, Jasjit S Suri, and Ayman El-Baz. Medical image segmentation: a brief survey. In *Multi Modality State-of-the-Art Medical Image Segmentation and Registration Methodologies*, pages 1–39. Springer, 2011.

[103] Marek Kowal, Paweł Filipczuk, Andrzej Obuchowicz, Józef Korbicz, and Roman Monczak. Computer-aided diagnosis of breast cancer based on fine needle biopsy microscopic images. *Computers in Biology and Medicine*, 43(10):1563–1572, 2013.

[104] TW Ridler and S Calvard. Picture thresholding using an iterative selection method. *IEEE Trans Syst Man Cybern*, 8(8):630–632, 1978.

[105] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1988.

[106] Chenyang Xu and Jerry L Prince. Generalized gradient vector flow external forces for active contours. *Signal Processing*, 71(2):131–139, 1998.

[107] Jihene Malek, Abderrahim Sebri, Souhir Mabrouk, Kholdoun Torki, and Rached Tourki. Automated breast cancer diagnosis based on gvf-snake segmentation, wavelet features extraction and fuzzy classification. *Journal of Signal Processing Systems*, 55(1-3):49–66, 2009.

[108] Norberto Malpica, Carlos Ortiz De Solórzano, Juan José Vaquero, Andrés Santos, Isabel Vallcorba, José Miguel García-Sagredo, and Francisco Del Pozo. Applying watershed algorithms to the segmentation of clustered nuclei. *Cytometry: The Journal of the International Society for Analytical Cytology*, 28(4):289–297, 1997.

[109] Sergey Ya Proskuryakov, Anatoli G Konoplyannikov, and Vladimir L Gabai. Necrosis: a specific form of programmed cell death? *Experimental Cell Research*, 283(1):1–16, 2003.

[110] A Alvarez, J Lacalle, ML Cañavate, D Alonso-Alconada, I Lara-Celador, FJ Alvarez, and E Hilario. Cell death. a comprehensive approximation. necrosis. *Microscopy. Science., Technology., Applications and Education*, pages 1017–1024, 2010.

[111] S Rello, JC Stockert, V l Moreno, A Gamez, M Pacheco, A Juarranz, M Canete, and A Villanueva. Morphological criteria to distinguish cell death induced by apoptotic and necrotic treatments. *Apoptosis*, 10(1):201–208, 2005.

[112] G Kroemer, L Galluzzi, Peter Vandenabeele, J Abrams, ES Alnemri, EH Baehrecke, MV Blagosklonny, WS El-Deiry, P Golstein, DR Green, et al. Classification of cell death: recommendations of the nomenclature committee on cell death 2009. *Cell Death & Differentiation*, 16(1):3–11, 2009.

[113] Neelam Sood, Jitendra Singh Nigam, Poonam Yadav, Shivani Rewri, Ankit Sharma, Anita Omhare, and Jaya Malhotra. Comparative study of cytomorphological robinson's grading for breast carcinoma with modified bloom-richardson histopathological grading. *Pathology Research International*, 2013, 2013.

[114] Mitko Veta, Paul J Van Diest, Stefan M Willems, Haibo Wang, Anant Madabhushi, Angel Cruz-Roa, Fabio Gonzalez, et al. Assessment of algorithms for mitosis detection in breast cancer histopathology images. *Medical Image Analysis*, 20(1):237–248, 2015.

[115] grows How cancer starts and spreads. Canadian cancer society. `http://www.cancer.ca/en/cancer-information/cancer-101/what-is-cancer/how-cancer-starts-grows-and-spreads/?region=on`, 2019. [Online; accessed 2019].

[116] Humayun Irshad. Automated mitosis detection in histopathology using morphological and multi-channel statistics features. *Journal of Pathology Informatics*, 4, 2013.

[117] Steven M Debol, Michael W Stanley, Shawn Mallery, Elizabeth Sawinski, and Ricardo H Bardales. Glomus tumor of the stomach: Cytologic diagnosis by endoscopic ultrasound-guided fine-needle aspiration. *Diagnostic Cytopathology*, 28(6):316–321, 2003.

[118] Badr AbdullGaffar, Mohamed Osman Kamal, Manar Khalid, Ravirani Samuel, and Rafeea AlGhufli. Atypical bare nuclei in liquid-based cervical cytology and their significance. *Acta Cytologica*, 53(6):637–643, 2009.

[119] Heang-Ping Chan, Datong Wei, Mark A Helvie, Berkman Sahiner, Dorit D Adler, Mitchel M Goodsitt, and Nicolas Petrick. Computer-aided classification of mammographic masses and normal tissue: linear discriminant analysis in texture feature space. *Physics in Medicine And Biology*, 40(5):857, 1995.

[120] Victor Goldberg, Armando Manduca, Daniel L Ewert, John J Gisvold, and James F Greenleaf. Improvement in specificity of ultrasonography for diagnosis of breast tumors by means of artificial intelligence. *Medical Physics*, 19(6):1475–1481, 1992.

[121] Yuzheng Wu, Maryellen L Giger, Kunio Doi, Carl J Vyborny, Robert A Schmidt, and Charles E Metz. Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer. *Radiology*, 187(1):81–87, 1993.

[122] Joseph Y Lo, Jay A Baker, Phyllis J Kornguth, and Carey E Floyd. Computer-aided diagnosis of breast cancer: artificial neural network approach for optimized merging of mammographic features. *Academic Radiology*, 2(10):841–850, 1995.

[123] Giorgio Roffo. Feature selection techniques for classification: A widely applicable code library. *Arxiv Preprint Arxiv:1607.01327*, 2016.

[124] Chris Seiffert, Taghi M Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. Rusboost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 40(1):185–197, 2010.

[125] Muneera Alsaedi, Thomas Fevens, Adam Krzyżak, and Łukasz Jeleń. Cytological malignancy grading systems for fine needle aspiration biopsies of breast cancer. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 705–709, Kansas City, MO, USA, November 13-16, 2017.

[126] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.

[127] Rui Yan, Fei Ren, Zihao Wang, Lihua Wang, Tong Zhang, Yudong Liu, Xiaosong Rao, Chunhou Zheng, and Fa Zhang. Breast cancer histopathological image classification using a hybrid deep neural network. *Methods*, 2019.

[128] Baris Gecer, Selim Aksoy, Ezgi Mercan, Linda G. Shapiro, Donald L. Weaver, and Joann G. Elmore. Detection and classification of cancer in whole slide breast histopathology images using deep convolutional networks. *Pattern Recognition*, 84:345–356, 2018.

[129] B. Miselis, T. Fevens, A. Krzyżak, M. Kowal, and R. Monczak. Deep neural networks for breast cancer diagnosis: Fine needle biopsy scenario. In *21st Polish Conference on Biocybernetics and Biomedical Engineering (PCBBE)*, Zielona Góra, Poland, 25-27 September 2019.

[130] Yun Jiang, Li Chen, Hai Zhang, and Xiao Xiao. Breast cancer histopathological image classification using convolutional neural networks with small se-resnet module. In *PloS one*, 2019.

[131] Seokmin Han, Ho-Kyung Kang, Ja-Yeon Jeong, Moon ho Park, Wonsik Kim, Won-Chul Bang, and Yeong-Kyeong Seong. A deep learning framework for supporting the classification of breast lesions in ultrasound images. *Physics in Medicine and Biology*, 62(19):7714–7728, 2017.

[132] Mehmet Ufuk Dalmış, Suzan Vreemann, Thijs Kooi, Ritse M Mann, Nico Karssemeijer, and Albert Gubern-Mérida. Fully automated detection of breast cancer in screening mri using convolutional neural networks. *Journal of Medical Imaging*, 5(1):014502 (9 pages), 2018.

[133] Thijs Kooi, Geert J. S. Litjens, Bram van Ginneken, Albert Gubern-Mérida, Clara I. Sánchez, Ritse M Mann, Ard den Heeten, and Nico Karssemeijer. Large scale deep learning for computer aided detection of mammographic lesions. *Medical Image Analysis*, 35:303–312, 2017.

[134] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1717–1724, 2014.

[135] ImageNet. `www.image-net.org`. Online.

[136] N. Bayramoglu, J. Kannala, and J. Heikkilä. Deep learning for magnification independent breast cancer histopathology image classification. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2440–2445, Dec 2016.

[137] Fabio A. Spanhol, Luiz Eduardo Soares de Oliveira, Caroline Petitjean, and Laurent Heutte. A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, 63:1455–1462, 2016.

[138] Samira Pouyanfar, Yudong Tao, Anup Mohan, Haiman Tian, Ahmed S Kaseb, Kent Gauen, Ryan Dailey, Sarah Aghajanzadeh, Yung-Hsiang Lu, Shu-Ching Chen, et al. Dynamic sampling

in convolutional neural networks for imbalanced data classification. In *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*, pages 112–117. IEEE, 2018.

[139] Sebastien C Wong, Adam Gatt, Victor Stamatescu, and Mark D McDonnell. Understanding data augmentation for classification: when to warp? In *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–6. IEEE, 2016.

[140] Ning Chen, Jun Zhu, Fei Xia, and Bo Zhang. Generalized relational topic models with data augmentation. In *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

Appendix A

Dataset of FNA biopsies

Five sets of data were being used to test the proposed computer-aided cytological grading systems (CGSs). The dataset contains FNA biopsy images collected during examinations from the Department of Pathology and Oncological Cytology of the Medical University of Wrocław, Poland. The images belong to two classes of cancer malignancy, namely intermediate (G2) and high (G3) malignancy grades. The lack of low malignancy (G1) images is caused by the fact that these cases very rarely require FNA only a few cases were reported in recent years at the Medical University of Wrocław. Preparation of the slides includes staining with Haematoxylin and Eosin. Regions of interest on the slides were digitalized with a resolution of 96 dots per inch (dpi) and a size of 764x572 pixels. Each dataset consists of images obtained with two different magnifications (100x and 400x) of the same tissue region for each patient. The 100x low magnification images are obtained by multiplying the 4x value of object power by the 10x value of eyepiece power. Whereas, the 400x high magnification images are obtained by multiplying the 40x value of object power by the 10x value of eyepiece power (see Fig. 62).

This pair of images (100x and 400x) describes a single case used for malignancy determination for a specific patient. The malignancy grading for all the patients was histopathologically validated using surgical biopsies graded using BR grading [19]. Each of the five datasets divided into two subsets depending on the malignancy grades (G2 and G3) of the samples. Further, for each sample, there is two magnification of the slides (100x and 400x) as described below for each of the used dataset.

Figure 62: The estimation of total magnification of cytological medical images used by pathologists taken from .

A.1 JELEN08 dataset

The first set labeled JELEN08, as used by Jeleń *et al.* [83], was taken from 22 patients and includes 14 patients with grade G2 (comprising 33 cases of pairs of 100x and 400x images) and 8 patients with grade G3 (comprising 13 cases of pairs 100x and 400x images). The figures 63–66 examples of the images belong to G2 and G3 cases with the low and high magnifications from JELEN08 dataset.

A.2 JELEN16 Dataset

The second set named JELEN16, as used in Jeleń *et al.* [84], is taken from 41 patients and includes 34 patients with grade G2 (comprising 100 cases of pairs of 100x and 400x images) and 7 patients with grade G3 (comprising 20 cases of pairs of 100x and 400x images). The original resolution of the images of this database was 2070x1548 pixels at 200 dpi. Thus, the size of the images was resized

to 764x572 pixels at 96 dpi to normalize the size of all the images in the datasets used. The figures 67–70 examples of the images belong to G2 and G3 cases with the low and high magnifications from JELEN16 dataset.

A.3 JELEN18 Dataset

The third set named JELEN18 is taken from 41 patients and includes 34 patients with grade G2 (comprising 100 cases of pairs of 100x and 400x images) and 7 patients with grade G3 (comprising 20 cases of pairs of 100x and 400x images). The original resolution of the images of this database also was 2070x1548 pixels at 200 dpi. Thus, the size of the images was resized to 764x572 pixels at 96 dpi as the size of the previous two datasets with the aim of normalizing the size of all the images in the datasets used. The figures 71–74 examples of the images belongs to G2 and G3 cases with the low and high magnifications from JELEN18 dataset.

A.4 JELEN_MERGE01 and JELEN_MERGE02 Datasets

The fourth dataset labeled JELEN_MERGE01 combines the JELEN08 and JELEN16 sets into a single dataset with a total of 63 patients. Whereas, the fifth dataset named JELEN_MERGE02 combines the JELEN08, JELEN16, and JELEN18 sets into a single dataset with a total of 104 patients.

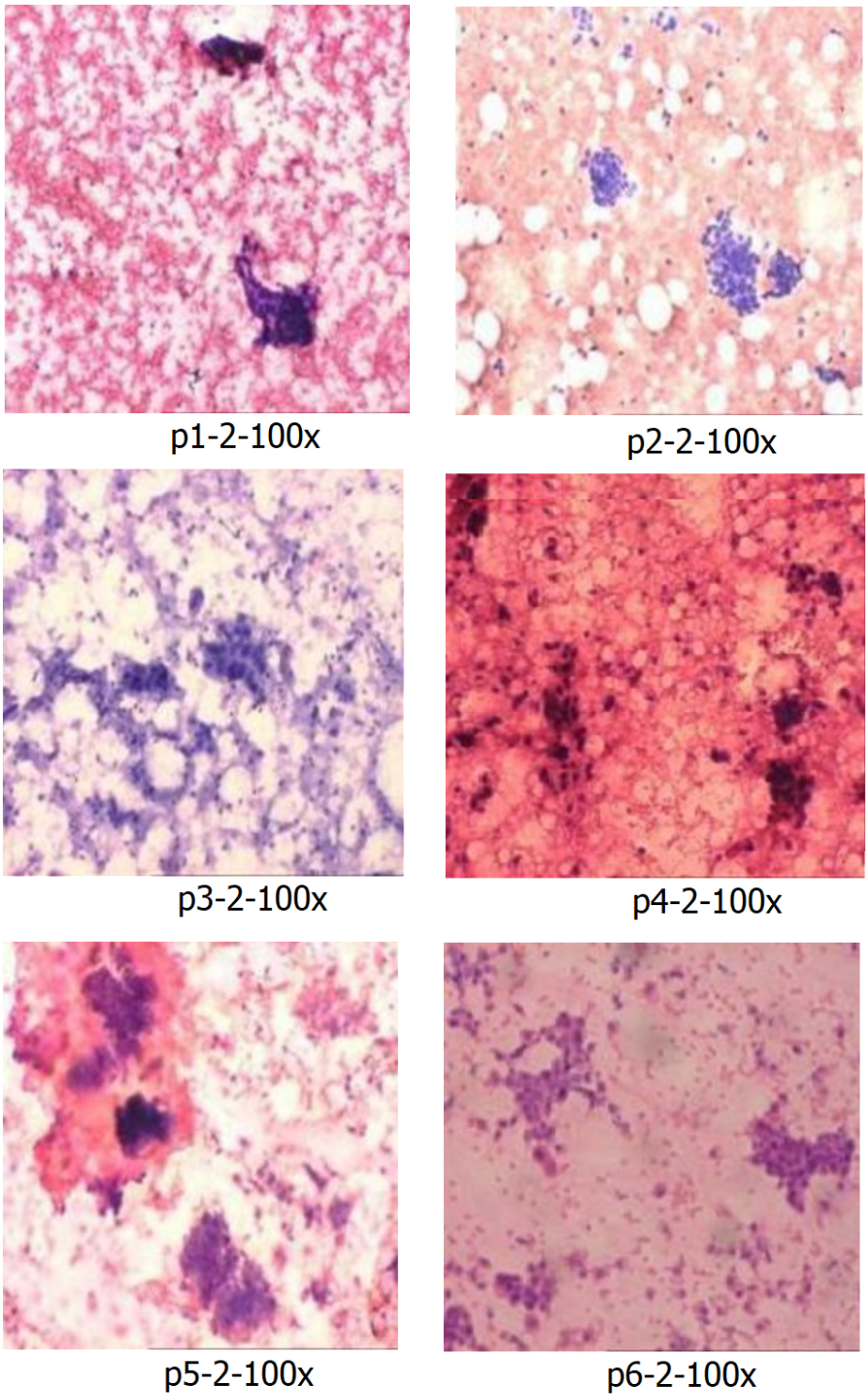


Figure 63: Example of low magnification images of G2 cases from JELEN08 dataset.

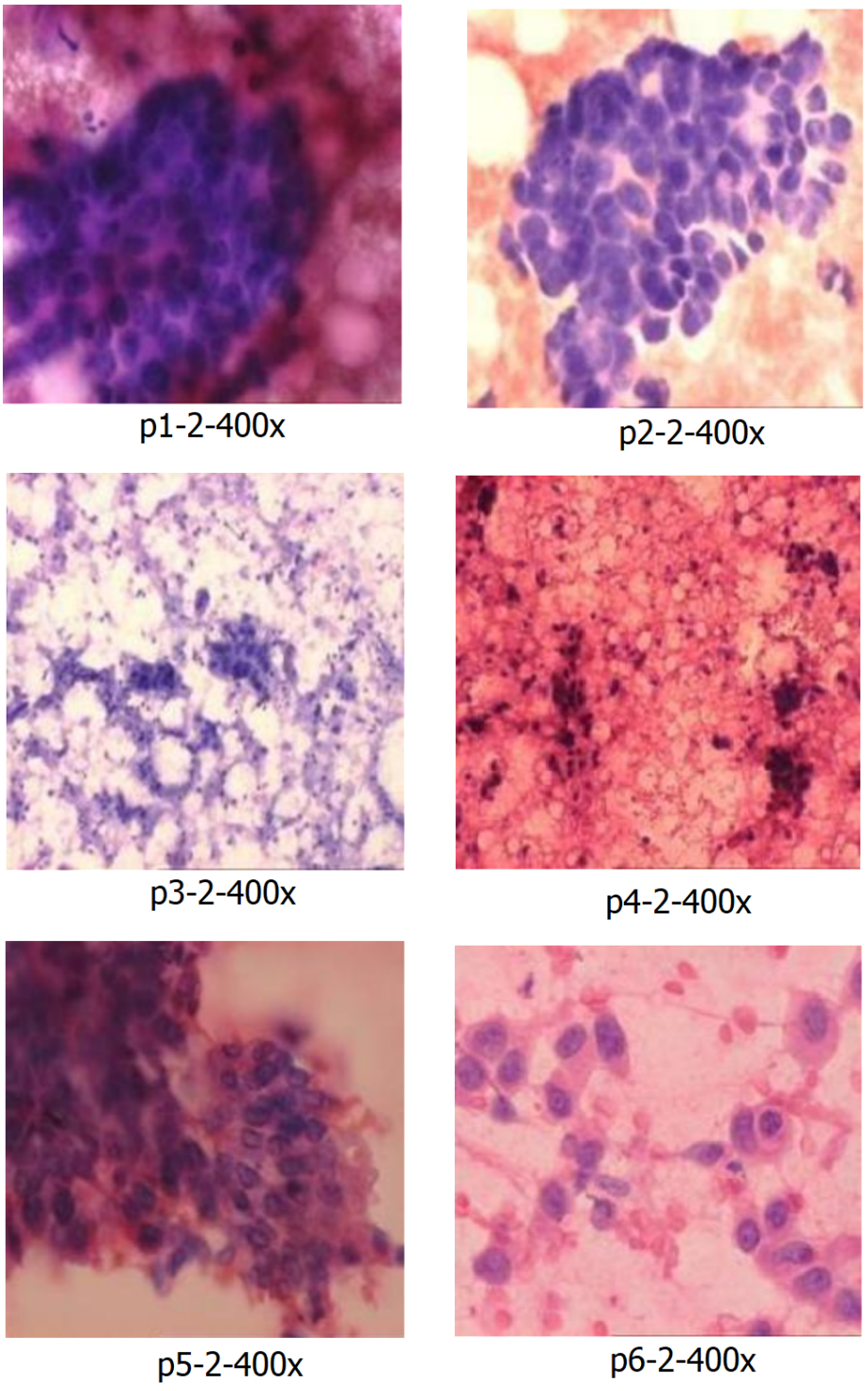


Figure 64: Example of high magnification images of G2 cases from JELEN08 dataset.

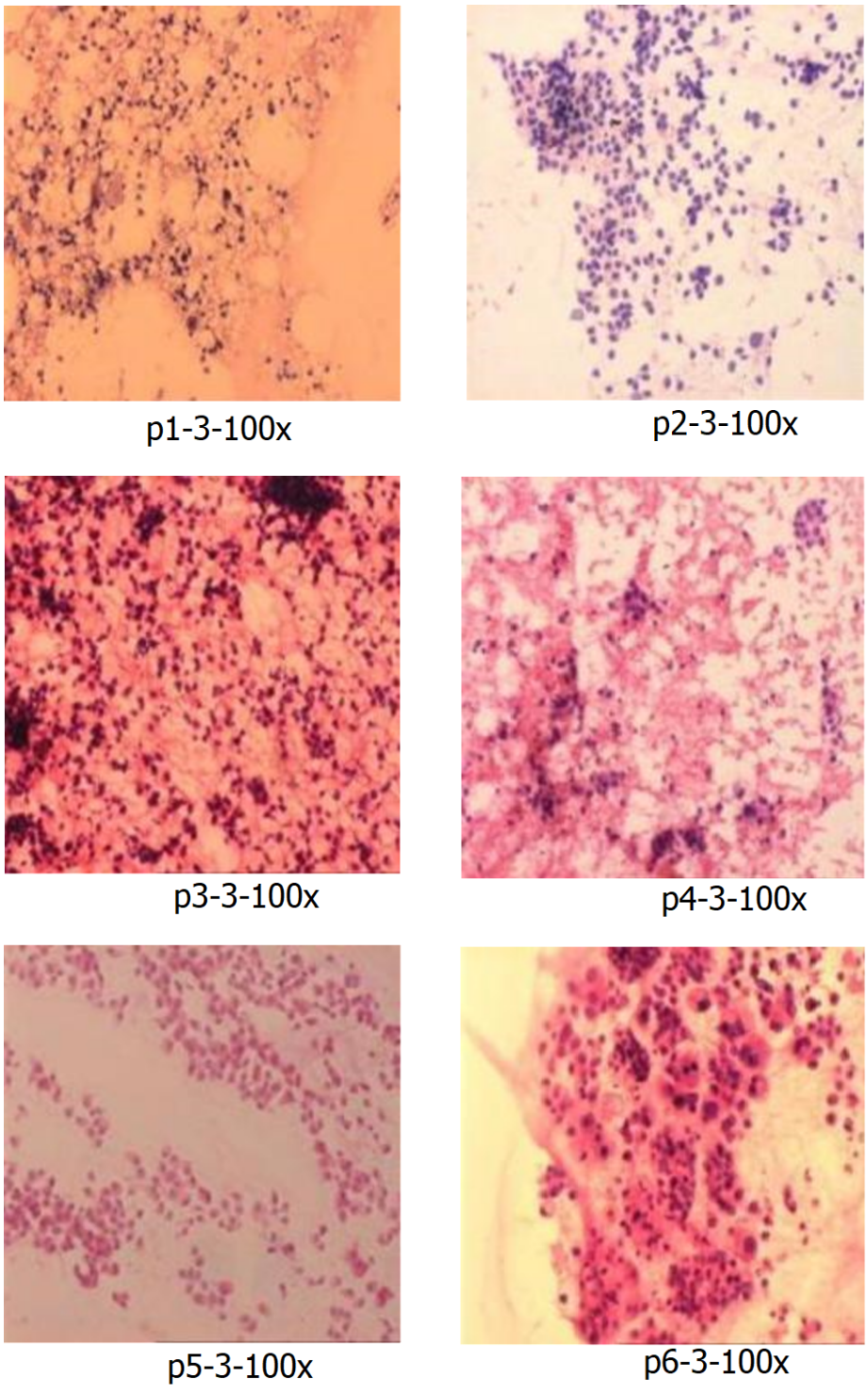


Figure 65: Example of low magnification images of G3 cases from JELEN08 dataset.

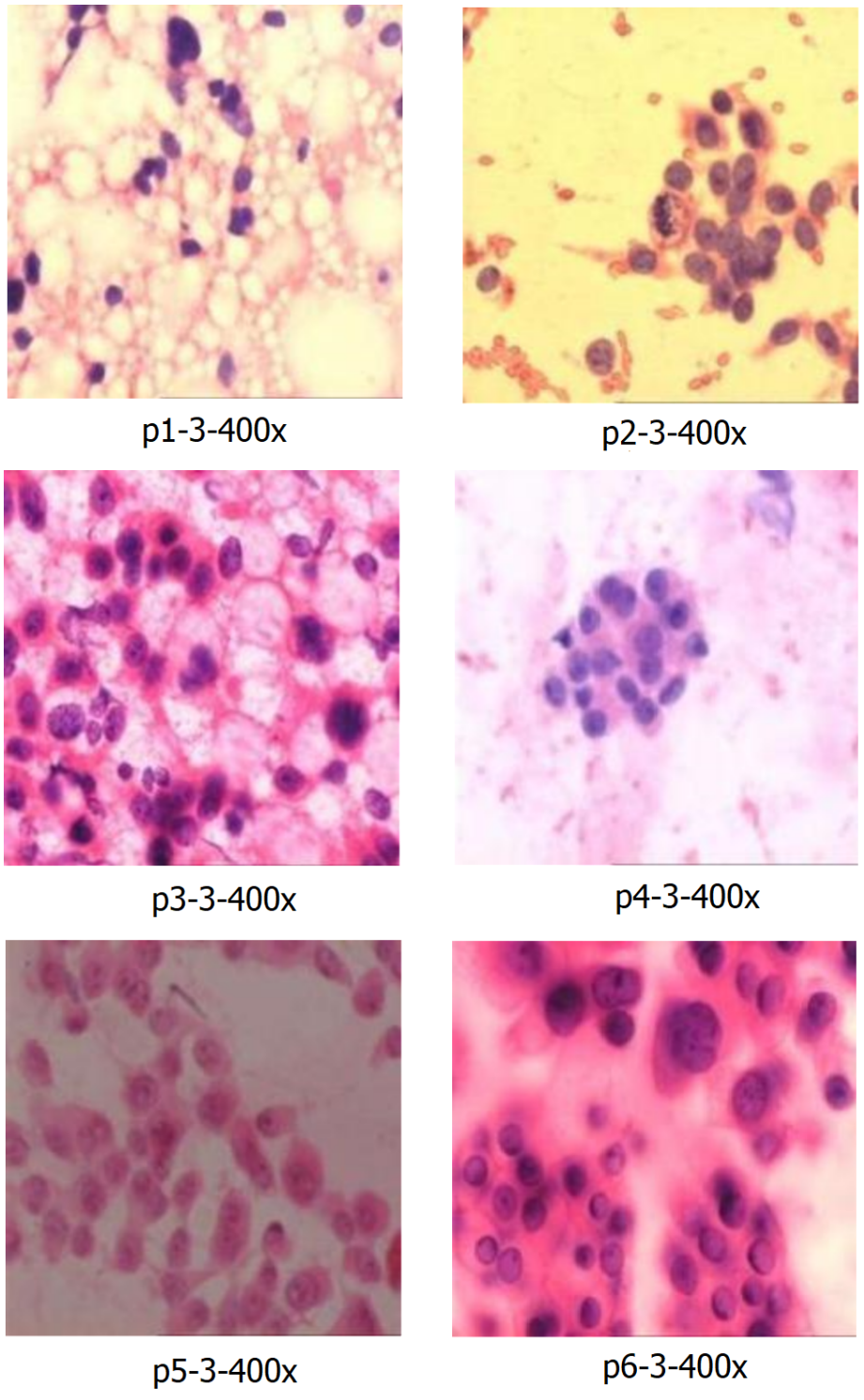


Figure 66: Example of high magnification images of G3 cases from JELEN08 dataset.

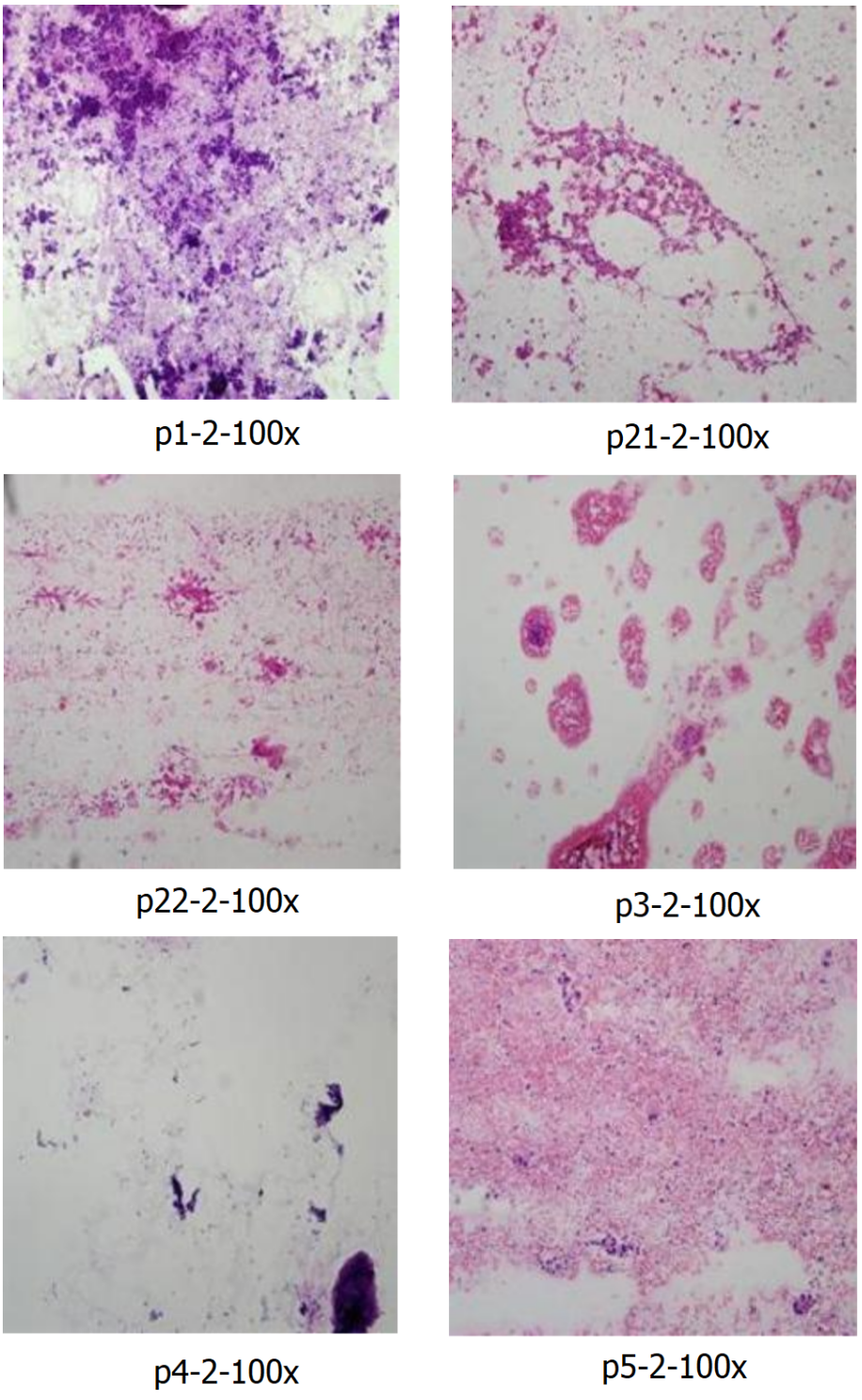


Figure 67: Example of low magnification images of G2 cases from JELEN16 dataset.

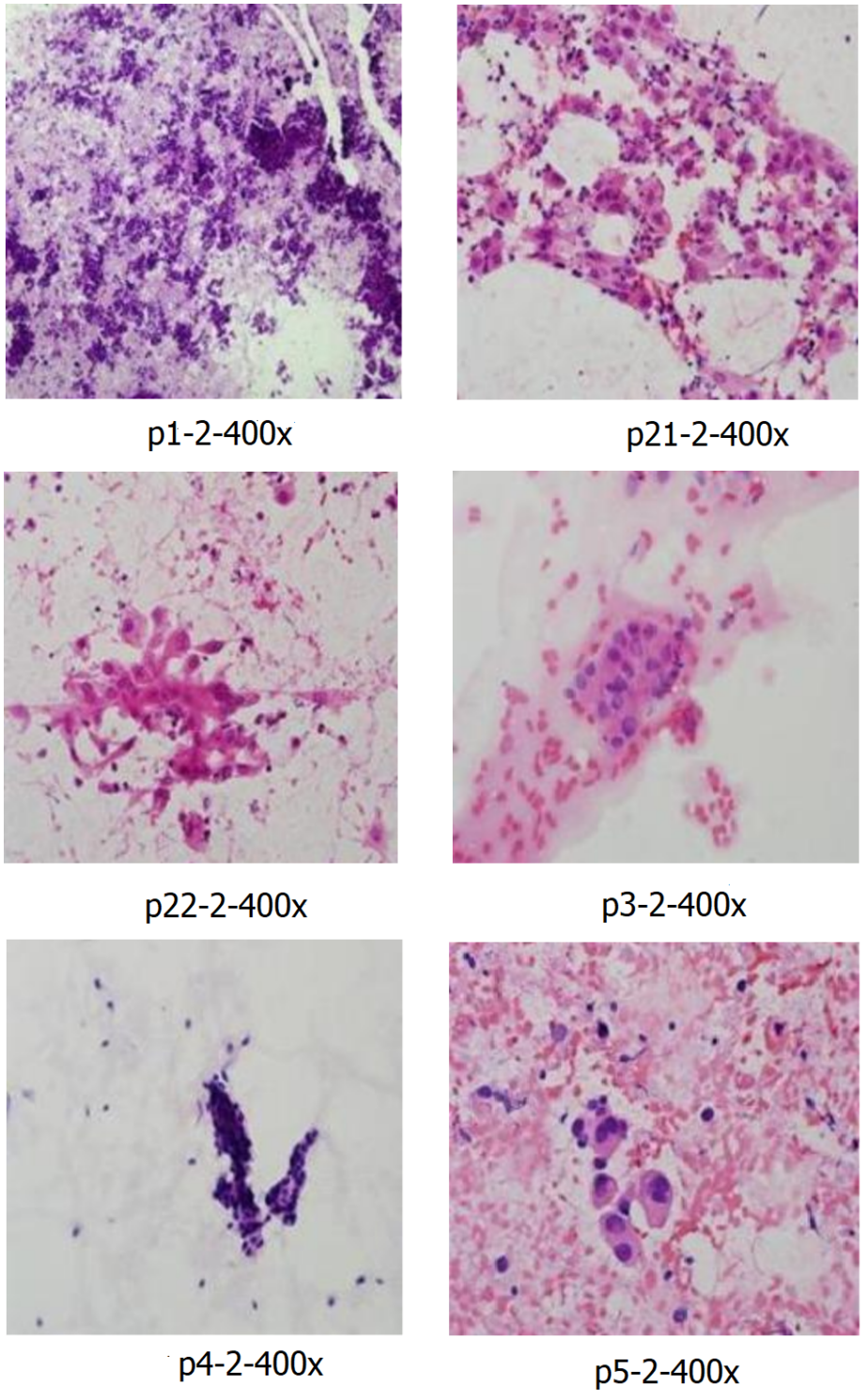


Figure 68: Example of high magnification images of G2 cases from JELEN16 dataset.

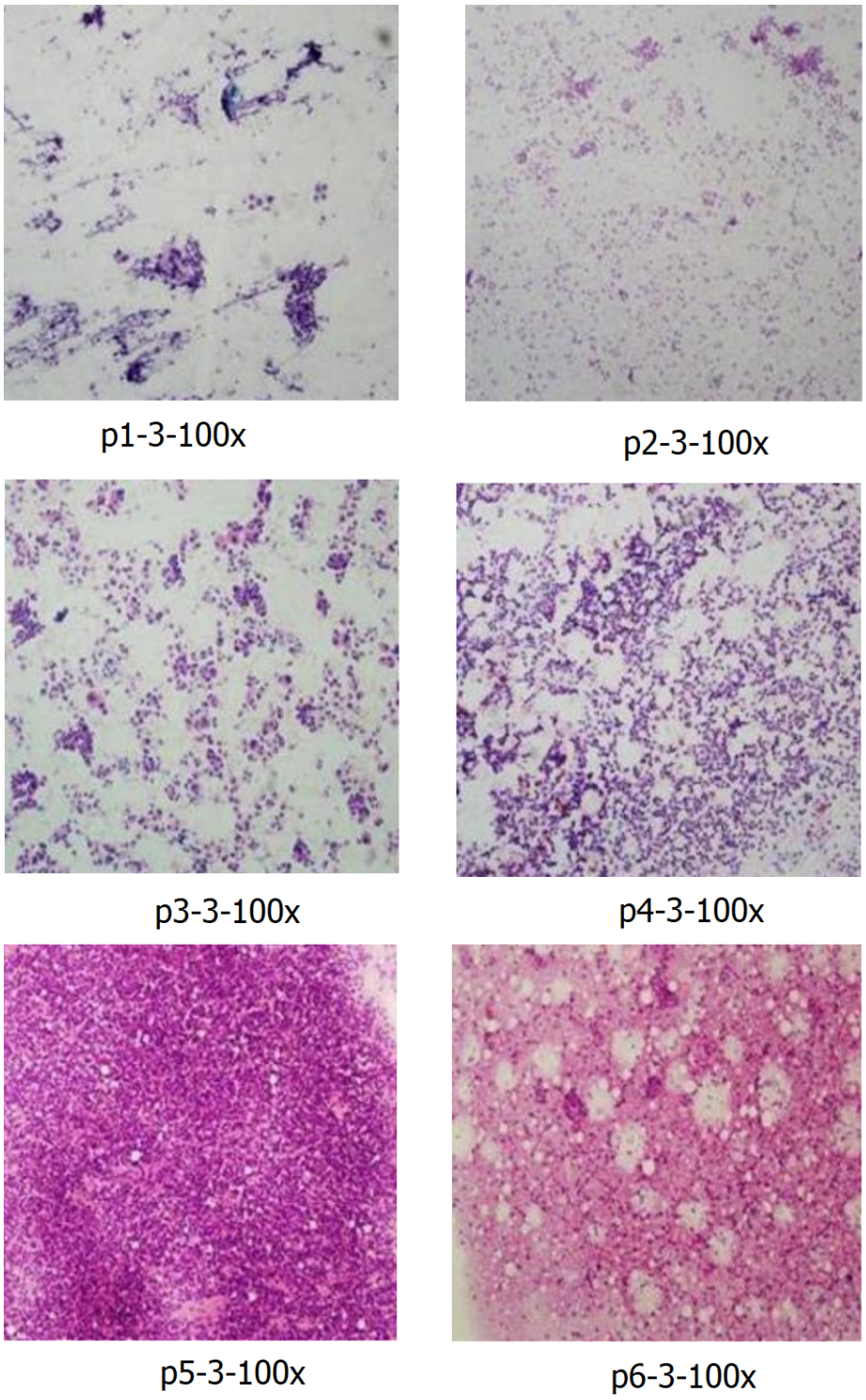


Figure 69: Example of low magnification images of G3 cases from JELEN16 dataset.

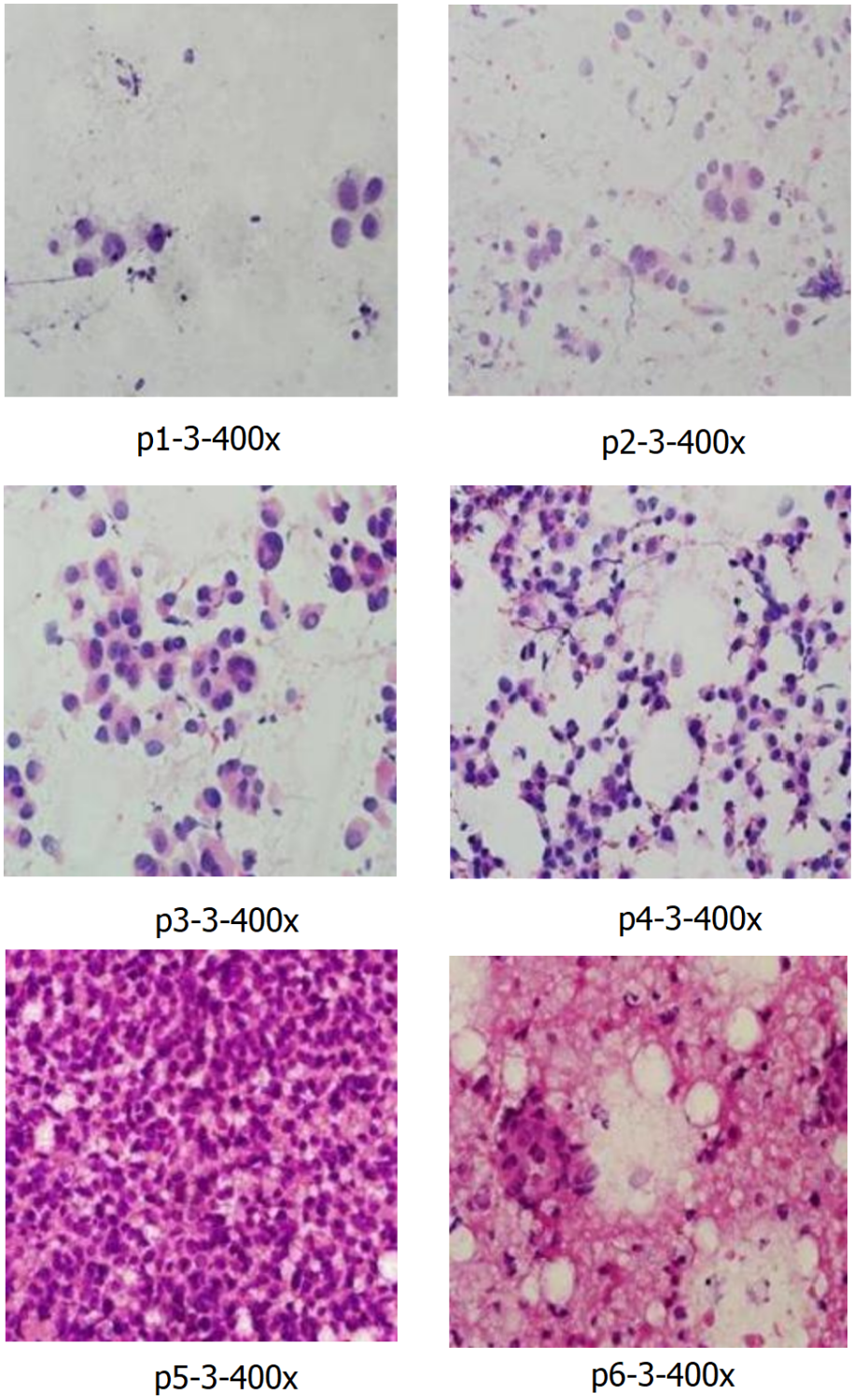


Figure 70: Example of high magnification images of G3 cases from JELEN16 dataset.

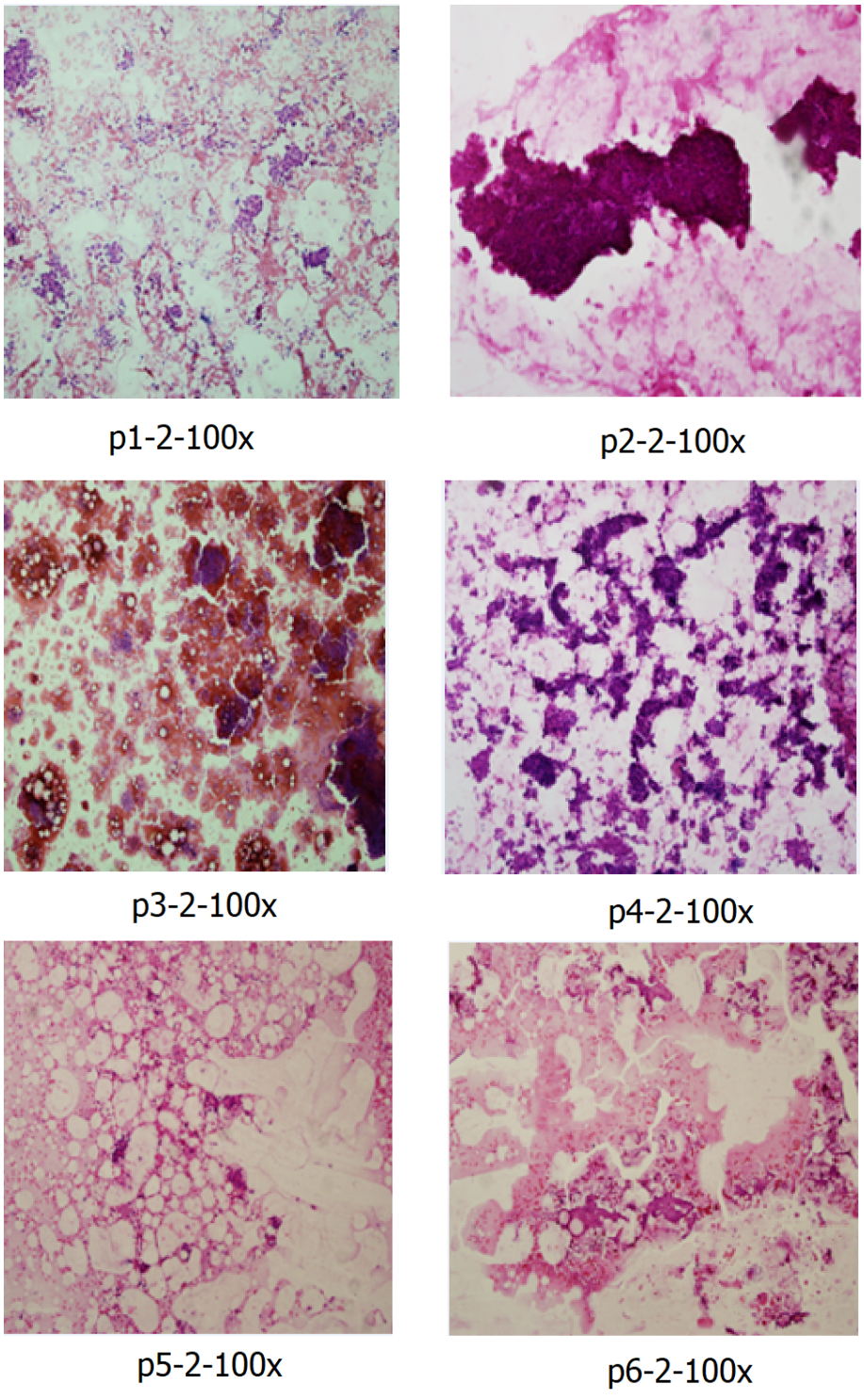


Figure 71: Example of low magnification images of G2 cases from JELEN18 dataset.

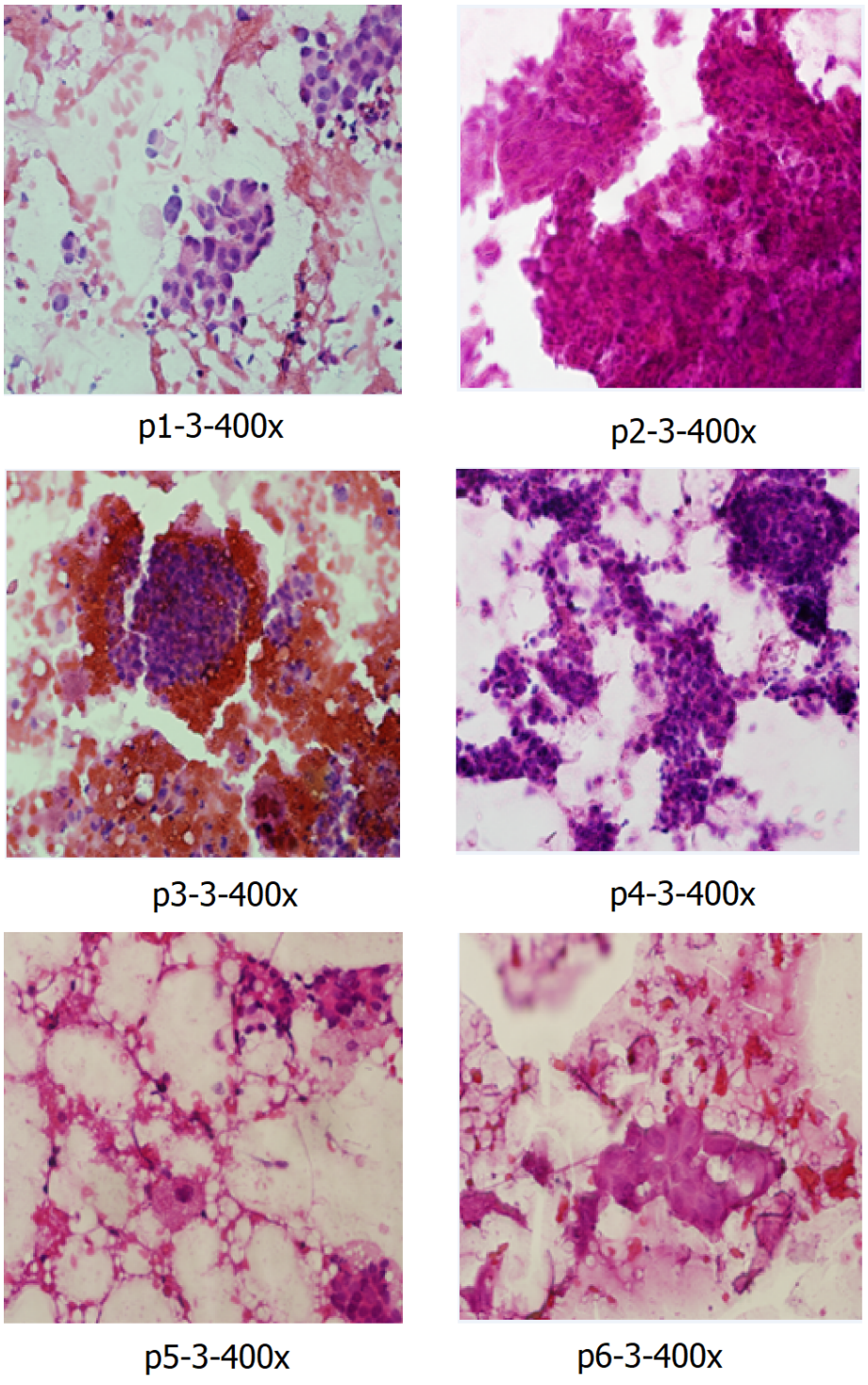


Figure 72: Example of high magnification images of G2 cases from JELEN18 dataset.

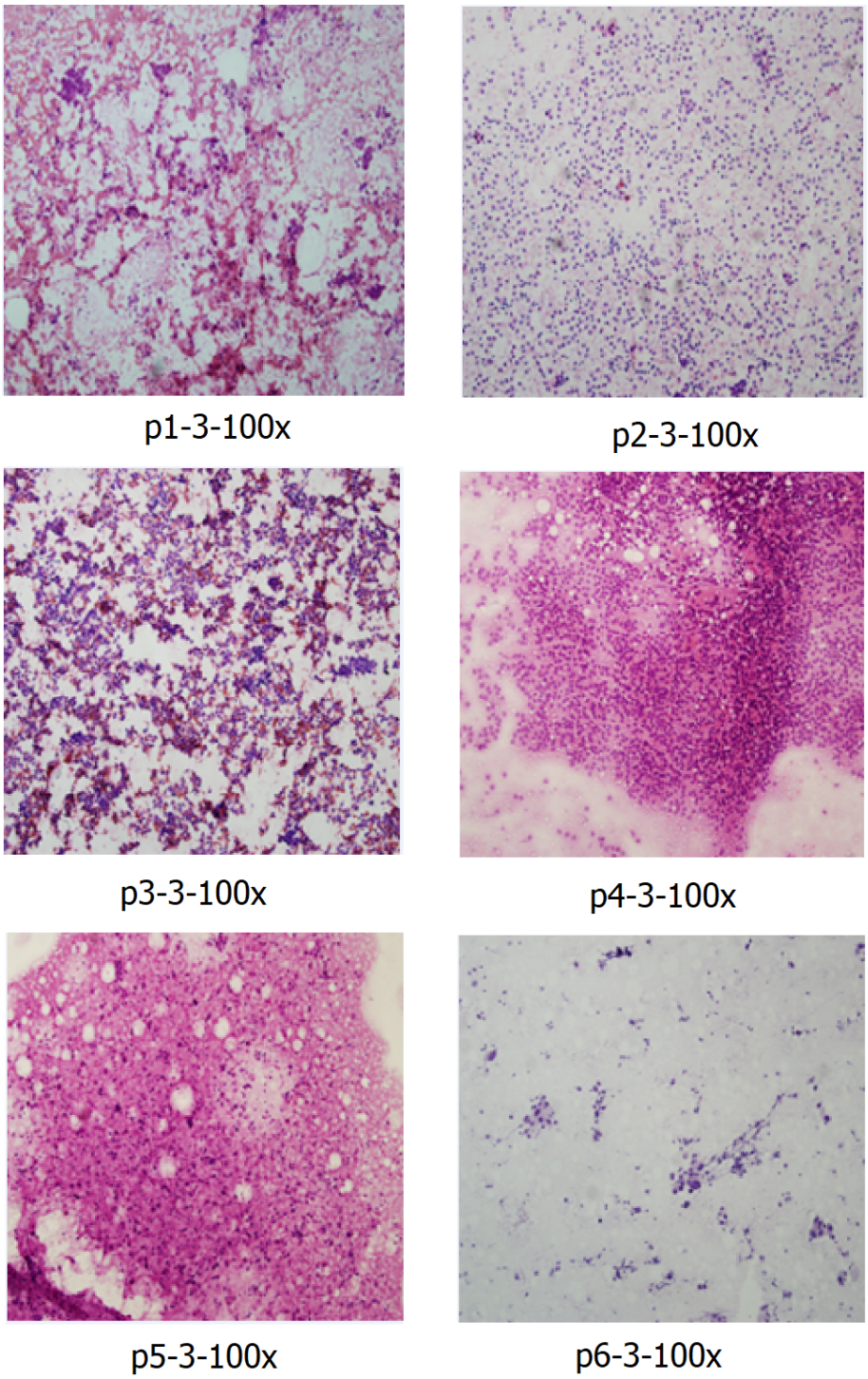


Figure 73: Example of low magnification images of G3 cases from JELEN18 dataset.

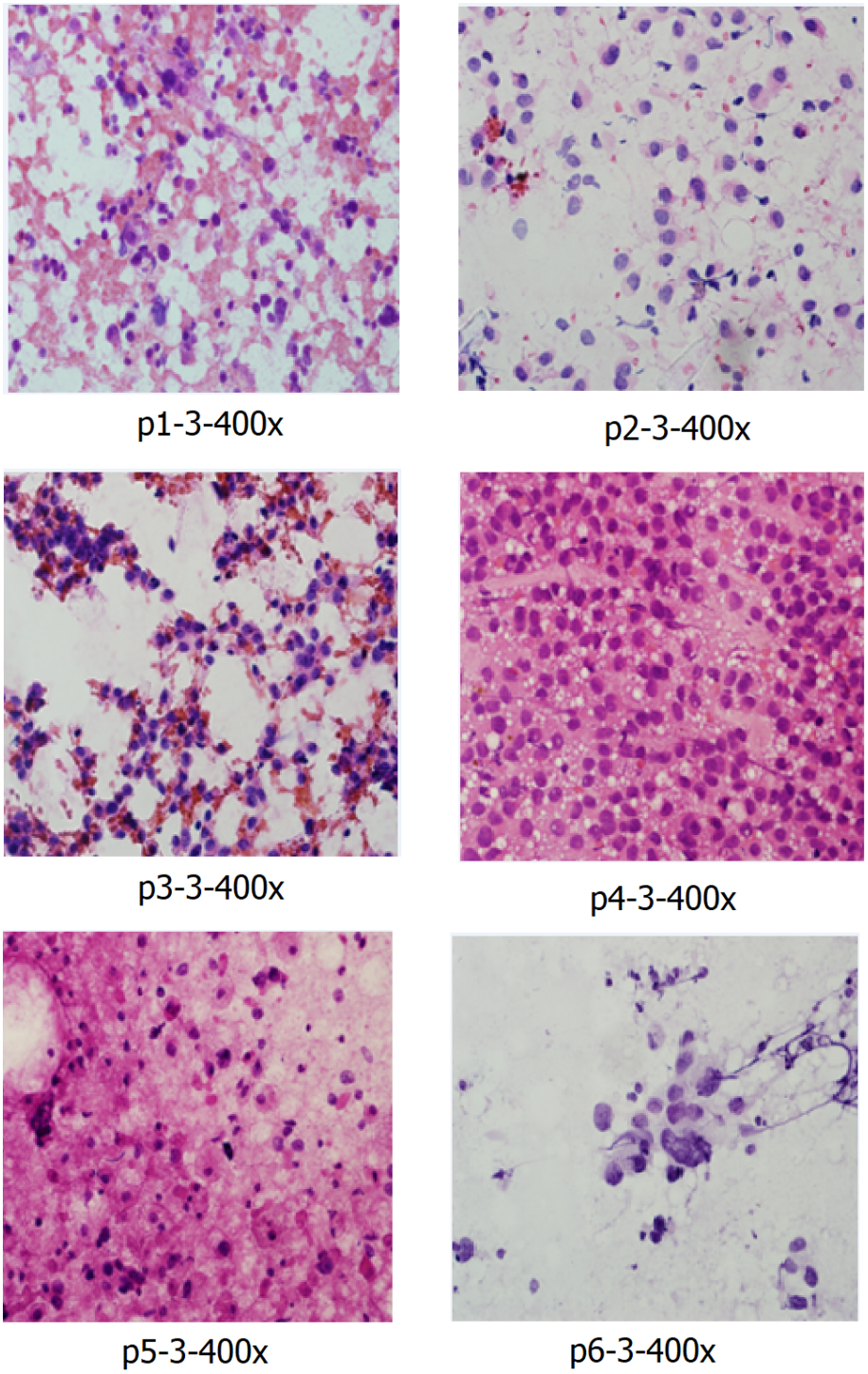


Figure 74: Example of high magnification images of G3 cases from JELEN18 dataset.